DSIE'15
Doctoral Symposium in Informatics Engineering

www.fe.up.pt/dsie15

Proceedings of the 10th Doctoral Symposium in Informatics Engineering

January, 29th and 30th, 2015
Porto, Portugal

Editors:
A. Augusto de Sousa
Eugénio Oliveira

Sponsors

U.PORTO

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

DEI Departamento de Engenharia Informática

DELTA CAFÉS

unicer

COPYRIGHT

1ª Edição/ 1st Edition 2015
ISBN: 978-972-752-173-9
Editors: A. Augusto Sousa and Eugénio Oliveira
Proceedings Design: Carlos Alex Sander Juvêncio Gulo
Graphical Design/Website: Pedro Leitão and Thiago Rúbio
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto

DSIE'15 SECRETARIAT:
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
Telephone: +351 22 508 21 34
Fax: +351 22 508 14 43

E-mail: dsie15@fe.up.pt
Symposium Website: http://www.fe.up.pt/dsie15

FOREWORD

STEERING COMMITTEE

Although living hard times in Europe, we believe that hope always come from wise people and more knowledge. PhD students organizing and collaborating with this event deserve our admiration for their efforts to keep alive their quest for new knowledge and better intellectual and technical skills. These are also times for strong wills guiding their way ahead.

2015 Doctoral Symposium in Informatics Engineering - DSIE'15 consubstantiates the 10th edition of a scientific meeting mainly organized by PhD students of the FEUP Doctoral Program in Informatics Engineering (PRODEI).

DSIE meetings have been held since the scholar year 2005/06 and the main goal has always been to provide a forum for discussion on, and demonstration of, the practical application of a variety of scientific research issues, particularly in the context of information technology, computer science and computer engineering. DSIE symposium comes out as a natural conclusion of mandatory ProDEI course called "Methodologies for Scientific Research" (MSR) leading to a formal evaluation of the students learned competencies.

The aim of this specific course (MSR) is to give students the opportunity to learn the processes, methodologies and best practices related to scientific research, particularly in the referred areas, as well as to improve their own capability to produce adequate scientific texts. With a mixed format based on multidisciplinary seminars and tutorials, the course culminates with the realization of the DSIE meeting, seen as a kind of laboratory test for the concepts learned by students. In the scope of DSIE, students are expected to play various roles, such as authors of the articles, members of both scientific and organization committees, as well as reviewers, duly guided by senior lecturers and professors.

DSIE event is then seen as a "leitmotif" for the students to be exposed to all facets of a scientific meeting associated with outstanding research activities in the area. Although still at an embryonic stage, and despite some of the papers still lack of maturity, we already can find some interesting research work, competent surveys and interesting perspectives about future work. At this moment, it was not essential, nor even possible, for most of the students in the first semester of their PhD, to produce sound and deep research results. However, we hope that the basic requirements for publishing an acceptable scientific paper have been fulfilled.

Each year DSIE Proceedings include papers addressing different topics according to the current students' interest. This year, the tendency is on "Knowledge Discovery", mainly text and data mining, computer-based (serious) games and simulation. There are also papers on on-line forums for higher education, distributed computing and gesture recognition.

The complete DSIE'15 meeting encompasses a two days program that includes also two invited talks by an outstanding researcher in Creative Computing and a recent PhD graduate in Health Information Retrieval.

Professors responsible for ProDEI program current edition, are proud to participate in DSIE'15 meeting and would like to acknowledge all the students who have been deeply involved in the success of this event that, hopefully, will contribute for a better understanding of the themes that have been addressed during the referred course, the best scientific research methods and the good practices for writing scientific papers and conveying novel ideas.

Porto, January 2015

Eugénio Oliveira and Augusto Sousa (ProDEI)

FOREWORD

ORGANIZING AND SCIENTIFIC COMMITTEES

DSIE'15 Organization and Scientific Committees welcome you to the 10th Doctoral Symposium in Informatics Engineering, 2015. The main goal of a scientific event is to discuss, disseminate and create knowledge. Organizing this conference proved to be a challenging opportunity for us to achieve this goal. Regardless of its small size, it demanded our commitment and hard work but also delivered the proudness of seen the successful concretization of our plans. We take this knowledge for our future and believe that every person enrolled with the DSIE has improved its knowledge.

As chairs, we've accepted the mission to make the 10th edition of the DSIE a special event. With great honour we gave our best to organize the conference and deliver the quality of work referenced by this series of conferences. Not only in the organizational part, but mainly regarding the contribution to the science community. And it was only possible because of the effort of all students in the Doctoral Program in Informatics Engineering.

We would like to thank all the senior members of the Scientific Committee for their dedication and involvement in the DSIE'15. We would also like to thank the significant help of Sandra Reis, from the Informatics Engineering Department of FEUP, and to all the sponsors of the Doctoral Symposium of Informatics Engineering for their support and involvement to help DSIE'15 to be a reality.

And we also would like to thank all participants of DSIE'15.

Porto, January 2015

Thiago Reis (Organization Committee Chair)

Hugo Barbosa and Telmo Morais (Scientific Committee Chairs)

CONFERENCE COMMITTEES

STEERING COMMITTEE

A. Augusto Sousa

Eugénio Oliveira

ORGANIZING COMMITTEE CHAIR

Thiago Reis

Organizing COMMITTEE

Carlos Gulo

Elis Silva

Hugo Barbosa

João Ulisses

Luciano Moreira

Pedro Leitão

Shazia Tabassum

Telmo Morais

Thiago Rúbio

SCIENTIFIC COMMITTEE CO-CHAIRS

Hugo Barbosa

Telmo Morais

SENIOR SCIENTIFIC COMMITTEE

Ana Paiva

António Coelho

Francisco Vasques de Carvalho

João Cardoso

João Faria

João Ferreira

João Mendes Moreira

Pedro Strecht

Rosaldo Rossetti

Raul Vidal

Rui Maranhão de Abreu

Rui Rodrigues

Sérgio Sobral Nunes

JUNIOR SCIENTIFIC COMMITTEE

Carlos Gulo

Elis Silva

Hugo Barbosa

João Ulisses

Luciano Moreira

Pedro Leitão

Shazia Tabassum

Telmo Morais

Thiago Rúbio

SPONSORS

DSIE'15 – Doctoral Symposium in Informatics Engineering is sponsored by:

CONTENTS

INVITED SPEAKERS

Amílcar Cardoso 10

Carla Teixeira 11

SESSION 1 - INFORMATICS AND EDUCATION

Online forums in Higher Education: empowering female participation 13
Luciano Moreira

Context-based learning games for children with cerebral palsy: a prototype 25
Elis Regina Silva and Jorge Silva

SESSION 2 - KNOWLEDGE DISCOVERY

A Survey of Merging Decision Trees Data Mining Approaches 36
Pedro Strecht

A Review of recent progress in multi document summarization 48
Shazia Tabassum and Eugenio Oliveira

Text Mining Scientific Articles using the R Language 60
Carlos Gulo and Thiago Reis

Characterizing Developers' Rework on GitHub Open Source Projects 70
Thiago Reis and Carlos Gulo

SESSION 3 - RESEARCH ON PROGRAMMING

Analysis and Evaluation of gesture recognition using LeapMotion 83
Pedro Leitão

Survey on Frameworks for Distributed Computing 95
Telmo Morais

Procedural Generation of Maps and Narrative Inclusion for Video Games 106
João Ulisses, Ricardo Gonçalves and António Coelho

SESSION 4 - VIRTUAL SIMULATION

A Multi-player Approach in Serious Games: Testing Pedestrian Fire Evacuation Scenarios 120
Marcos André Oliveira, Nelson Miguel Pereira, Joao Emílio Almeida, Rosaldo Rossetti and Eugénio Costa Oliveira

3D Simulation Environment: Education and Training 132
Hugo Barbosa

INVITED SPEAKERS

INVITED SPEAKER

AMÍLCAR CARDOSO

Prof. F. Amílcar Cardoso is a Full Professor at the Department of Informatics Engineering of the University of Coimbra. He developed pioneering work on Computational Creativity in the 90's, and assumed since then an active role in the area.

He currently coordinates the Doctoral Plan on Sciences and Technologies of Information of the Univ. Coimbra, and also the Cognitive and Media Systems Group of CISUC, a team with 17 PhDs that performs research on Artificial Intelligence, Media Systems, Ubiquitous Systems and ICT for Education. He is currently involved in two EU projects on Computational Creativity: the FET CA PROSECCO (Promoting the Scientific Exploration of Computational Creativity) and the FET/ICT ConCreTe (Concept Creation Technology).

INVITED SPEAKER

CARLA TEIXEIRA LOPES

Carla Teixeira Lopes is graduated in Informatics and Computing Engineering from University of Porto. From the same university she also received an MSc degree in Information Management and a PhD in Informatics Engineering.

After 5 years as a lecturer at the School of Allied Health Sciences of the Polytechnic Institute of Porto, she is, since 2008 an assistant professor in the Department of Informatics Engineering, University of Porto. Her research interests lie at the intersection of information retrieval and human-computer interaction.

She is interested in studying information search behaviour and in developing tools that help people search more successfully. Lately, she has been focused in exploring how context can help improve the experience of health consumers searching the Web.

SESSION 1

INFORMATICS AND EDUCATION

Online forums in Higher Education: empowering female participation
Luciano Moreira

Context-based learning games for children with cerebral palsy: a prototype
Elis Regina Silva and Jorge Silva

Online forums in Higher Education: empowering female participation

Luciano Moreira

Faculdade de Engenharia da Universidade do Porto, Portugal
lucianomoreira@fe.up.pt

Abstract. Forums are widely available tools in educational platforms, but there is no consensus on their benefits. This paper is aimed at investigating students attitudes towards online forums, the content contributions published and if forums promoted a more equal participation between male and female students. The final evaluative syntheses of the students (n=55) and the posts published in one forum (n=49) were analysed. The results indicated that the students had positive attitudes towards forums, with male students being more critical. The content contributions consisted mainly of reflections and affirmations. Female students were as actively engaged in discussions as male students, increasing their participation when compared with the practical lessons. This study is limited by its sample size and by the fact that only one analyst coded the material. Further investigation is required to know if male and female students have different patterns of publishing.

Keywords: Forums, Higher Education, digital media, sex inequality

1 Introduction

In the last decades, economically developed societies have gone through a migration process of many of their basic structures, including not only finances or bureaucracy, but also human communication and socialization. Eventually, education, including Higher Education institutions, also reflected and took part of this unfinished digital metamorphosis.

Digital and technology-enhanced learning (web-based learning) could be an ally to change the landscape of teaching and learning in Higher Education. Nonetheless, evidence shows that the integration of digital media in pedagogic settings still faces several obstacles [1, 2]. While technological issues or insufficient training are often said to be the very own causes of misuse or incipient use of digital tools in education, the picture seems to be more complex, involving psycho-sociological factors, such as attitudes and values [3]. Furthermore, the benefits of the digital media to the

effective empowerment of citizens (including students) are disputable, particularly when one things of minority social groups [4], *i.e.*, groups with less power, such as females.

Despite important progresses towards sex equality, there is still a gap in female participation in several social spheres, from politics to science. In fact, although the number of female students in Higher Education has been changing dramatically [5], recent studies confirm that it is still hard for them to have access to scientific careers, because of bias selection and gender stereotypes [6, 7]. Scientific production also mirrors the predominance of male researchers [8]. Not surprisingly, educational settings and classroom context reflects this trend, which is probably anchored in gender stereotypes, but also in classroom culture [9].

Despite of this, successful experiences do take place – involving both teachers and students – where communication is triggered and directed to educational discussions. This was the case of a course at the Faculty of Science of the University of Porto where virtual forums where used, getting as much participation as lectures and practical lessons. As time passed by, we observed that, in the context of lectures and practical lessons, male students engaged more actively in collective, oral discussion. This is a problem if one wants to empower all students as equal to participate in the political and scientific debates of contemporary society. Is it possible that female students participate more actively in the course's digital forums? By looking closely to students' attitudes and content contributions on the course forums, we were aimed at understanding if forums enabled female students to participate at least as actively as their male colleagues in the discussion.

The results of this research showed that students acknowledged the value of the forums as an alternative means to express themselves and know the others' perspective on relevant topics. The number and types of content contributions to the course's forums of female and male students were very similar and, as such, one can say that female students became more actively engaged in the digital forums. This research demonstrated that forums about relevant topics can be an important feature to include in the design of a Higher Education course, and, furthermore, that they seem to empower female participation. The question for future research is to understand if the digital participation transubstantiates into a more active engaged in non-virtual forums.

In the next section, we started by offering the reader a review of the relevant literature to investigate the relationship between the use of forums and sex inequality. In section 3, we described the research methods. In section 4, results were detailed exposed while in section 5 we discussed their meaning. In section 6, conclusions, limitations and future studies were presented.

2 Related work

Forums are widely present in online learning and blended-learning and are perceived as useful tools to use in Higher Education [10]. In fact, they are technological

tools that integrate most Learning Management Systems (LMS) (such as Moodle, Canvas or Blackboard) by default. As boundaries in Higher Education institutions become wider, forums have become the object of a growing interest by researchers [11].

In a recent survey of the Portuguese scientific production indexed in SCOPUS and Web of Knowledge on the topic of "online learning", Morais, Moreira and Paiva [12] showed that forums were an important source of data. For example, they provided researchers with facilitated access to students' or trainees' thoughts and ideas on several issues, such as training itself or general social topics; these thoughts and ideas were not directly enquired on but indirectly retrieved from users' entries (*i.e.*, posts published in forums) and were, for that reason, context-dependent.

LMS, such as Moodle, which incorporates forums, have been criticized as being old and ineffective tools with several limitations in environments that are increasingly open and connected [13] or, to use an expression borrowed from Jenkins [14], in a participatory culture. This is not the time or place to analyse this question, but one should not let it go without referring that the question is being asked in a naïve manner. In fact, as in this research we were focused on the students' attitudes towards forums, we aimed at getting new insights on what made the forums a successful tool in the context of this specific course.

Researchers have been looking at forums with a diversity of purposes. Some authors have been more interested in understanding interaction patterns [15]. Other authors have been more interested in understanding the development of collaborative work and what patterns emerge when they use text analysis [16], on facilitating online discussions [17] and knowledge building [18]. In this research, we were mainly interested in the type of contributions that students published in the forums.

Based on her pedagogical practice, Hughes [19] started to ask students to identify what type of content contributions they have written before they published in the course's forums (see Table 1). This proposal has not yet been tested, and, consequently, this research aimed at capturing its heuristic value to categorize students' posts.

The studies available do not solve the controversy on whether or not digital forums help equalizing the participation between sexes. For instance, His and Hoadley [20] supported the view that male students were more actively engaged in productive scientific discussion through electronic forums. Lim and Nahyun [21] also found that males seemed to appreciate more and have more chances to develop literacy skills than females by using Wikipedia. On the other hand, Prinsen, Volman and Terwel [9] showed that the picture is more complex. They reviewed a set of studies that focused the behavior of male and female students both on computer-mediated communication (CMC) and on computer-supported collaborative learning (CSCL) and found that there were differences in the degree and type of participation between sexes. Male students tended to be more actively engaged in CMC while in CSCL the situation were more balanced, although male students were more assertive in their statements and female more prone to agree. In another recent study, the same authors [22] found out that girls profit more than boys from participating in the elaboration of a programme in CSCL-environments, although this effect may be

connected with stereotyped patterns of communications (girls tend to ask and elaborate more than boys). Consequently, more research is necessary to understand if online forums can equalize male and female students and that is why it is important to study not only the degree but also the type of participation as we aimed at doing.

Table 1. Types of content contributions
(Hughes [19])

Category	Definition
a. Reflection	comments and initial thoughts, especially when asked to initially reflect about a course reading
b. Expansive Questions	about the content read or ideas posted by others; i.e., questions that spur others to think about or explore deeply the content or other ideas perhaps by using comparisons or metaphors
c. Substantive Insights	on the content read, including reinterpretations with the use of personal/ outside examples
d. Collegial Challenges	on the content read or ideas posted by others; i.e., disagreements with specific support for your perspective
e. Personal Realizations/Transformations	explicating significant shifts in your perspective on the topic due to the readings or the discussion
f. Clarification	about the question or comment posed – respondent feels clarification is needed before being able to respond
g. Affirmation	of ideas that participant has posed, including reasons why the respondent agrees or affirms the ideas.
h. Connection	to other peers' or the instructor's posts. You may direct readers to another post with similar ideas and content as your own or make a comment that connects or extends another person's ideas while explicitly acknowledging the connection.
i. Reiteration	when you are unsure of the meaning of a post, you may reiterate back to the poster what you think they mean, asking them to check if your interpretation correctly captures their intended meaning.
j. Summary/Wrap Up	often used by the instructor or moderator of a topic to summarize or bring together ideas across the discussion or part of the discussion.

In the next section, we will try to describe the methods used and the research questions that guided this research and that operationalized this problem within the specific context of a higher education course.

3 Method

This was a non-experimental, exploratory research, which used a qualitative methodological approach. In the following lines, we indicated the research questions,

described the context where the study took place and its participants, *i.e.*, students enrolled in the class, as well as data collection and data analysis procedures.

3.1 Research questions

Keeping in mind the empirical observations we have made on the students' participation and the scientific literature reviewed, we were not able to formalize clear hypotheses, but we were able to identify the following research questions:

1. What were the students' attitudes towards the course's forums?
2. What type of content contributions did male and female students published in the course's forums?
3. What was the heuristic value of Hughes [19] content contributions categories?
4. Did the forums empowered the participation of the female students?

3.2 Context and participants

This investigation occurred within the context of a Higher Education optative course on Personal and Professional Development at the Faculty of Sciences that took place in the 2nd semester of 2013/2014. Lectures consisted of a series of weekly seminars on topics such as employability, science and religion, scientific production, with invited speakers, followed by discussions. Practical lessons were supported by group dynamics and addressed personal and professional development topics, such as effective communication, group collaboration and time management. There were also virtual lessons that took place in the course's Moodle platform.

Students were required to participate in the forums discussing either the topics of the seminars or the topics of the virtual practical lessons. Each forum had a foreword written by the teachers. When the forums were about the seminars, students were asked to share their personal comments on the topics; whenever the forum was part of a virtual class, guidelines were given in a more detailed way and other resources were presented, such as further reading or videos. Students were informed that their participation in the forums was to be reflected in their final course grade.

Table 2 shows the distribution of students and final synthesis by gender and programme. As one can observe in Table 1, 83 (nearly 61%) students were female and 54 (nearly 39%) were male; the great majority of them were enrolled either in Biology (51 students, nearly 37%), Chemistry (41 students, nearly 30%) or Mathematics (28 students, nearly 20,4%), while the others were distributed by other programmes.

Table 2. Distribution of students by sex and programme

Variables		Students enrolled
		Frequency (%)
Sex	Female	83 (60,6%)
	Male	54 (39,4%)
Programme	Landscape Architecture	1 (0,7%)
	Biology	51 (37,2%)
	Biochemistry	2 (1,5%)
	Engineering Sciences	4 (2,9%)
	Earth and Environment Sciences	6 (4,4%)
	Geology	1 (0,7%)
	Mathematics	28 (20,4%)
	Chemistry	41 (29,9%)
	Mobility	3 (2,2%)
Total		137 (100%)

3.3 Data collection and data analysis

The data included (1) the personal syntheses of the students (no more than one page) that consisted in a reflection on their participation in the course and (2) the online forums of the course's Moodle platform.

A qualitative analysis has been conducted using NVIVO 10 for Windows, a qualitative data analysis software. Final syntheses and forums were retrieved from the course's Moodle platform and inserted into the software. After a preliminary and exhaustive reading, the *corpus* of analysis was constituted by means of a query that identified the content associated with forums using the following key-words: "fóruns OR forum OR fórum OR forums". We should remind the reader that the course was given in Portuguese.

Table 3 shows the initial material available and the *corpus* that was further analysed. Initial material consisted on the personal evaluative synthesis that each and every student was asked to write. The *corpus* that was further analysed only encompassed the syntheses that contained at least one reference to the forums. The context unit was the surrounding paragraph of the coding unit. Following Bardin [23] we define context unit as the unit that allows us to understand the meaning of the coding unit, in this case the theme. The coding unit is the segment of the *corpus* of analysis that is to be categorized and that ultimately can be counted. A theme is a nucleus of meaning that can be identified by the criteria specified by the analysts. In this case, we did not use *a priori* categories but instead we followed an approach inspired by the grounded-theory [24] according to which a constant revision method was used to reorganize categories as they emerged from the analysis. The categories used were attitudes, *i.e.*, "a psychological tendency, that is expressed by evaluating a particular entity with some degree of favor or disfavor" [25, p.269]. The *corpus* included 35 (63,6%) final syntheses from female students and 20 (36,4%) from male students.

Table 3. Material available and *corpus* of analysis

Variables		Material	*Corpus*
		Frequency (%)	Frequency (%)
Sex	Female	70 (64,2%)	35 (63,6%)
	Male	39 (35,8%)	20 (36,4%)
Total		109 (100%)	55 (100%)

Aside forums that were used to discuss doubts or to deliver group works, the course had 11 forums. As this is an exploratory study, we selected only the forum that had more visualizations (3041 views) to analyse. At this forum, 49 posts were analysed (19 new entries and 30 replies or threads). The context unit was the thread or, if necessary, the group of threads that followed a new entry, while the coding unit was the type of content contributions according to Hughes' proposal [19] (*vide* Table 1 in section 2).

4 Results

As we mentioned in the previous section, forums were highly participated. Students not only published their forums as they read other students' contributions. In this section, we presented the results obtained through the qualitative analysis. In the first place, we were aimed at mapping the attitudes of the students towards the course's forums and eventually to verify if we could find any sex-based effect.

Table 4 shows the attitudes towards forums (total and by sex) that were coded in the personal evaluative syntheses. We coded every segment of text that expressed a positive or negative view towards the forums as an attitude (coding unit). In the Table 4 we indicated the number of sources where a specific attitude was found.

As one can observe, there are more total sources coded with positive attitudes than sources coded with negative attitudes. Not only is this true, as when we look closely to the negative attitudes we can see that the majority was connected with the perceived participation, *i.e.,* specific references that the students made on their own participation. Sometimes, students, admitted that they could have been even more actively engaged in the discussions. A few students criticized the eventual redundancy of the topics, they showed surprise for they did not expect that their participation in the forums was so hard, they wished the discussions were more controversial and finally one student simply did not like this modality of communication.

On the other hand, positive attitudes did not refer only to the strong participation or commitment with the activities but they also enlightened the relevance and novelty of the themes proposed by the teachers and they explicitly referred a few personal development benefits that they took from their participation: dialogue and meeting other's perspective, self-expression and improvement of their writing skills.

Table 4. Attitudes towards forums

Evaluation	Attitudes	Male students (Sources coded)	Female students (Sources coded)	Total (Sources coded)
Positive	Participation	6	10	16
	Relevance	3	12	15
	Dialogue/Meeting others' perspectives	3	10	13
	Self-Expression	4	4	8
	Improving writing skills	2	3	5
	Total	12	27	39
Negative	Less participation	7	6	13
	Scheduling of post	2	1	3
	Redundancy	0	1	1
	Work	0	1	1
	Uninteresting	0	1	1
	Lack of contradictory	1	0	1
	Total	10	10	20

Female students showed more favourable attitudes than unfavourable attitudes towards forums. This is not surprising because 64,2% of the material analysed was written by female students who were also in majority in the course. It is, however, interesting that this proportionality was not mirrored in the negative attitudes towards forums.

Table 5 shows the students' contributions to the forums. Both female and male students published new entries and replies to previous posts in the same degree.

Table 5. Contributions

Type of entry	Male students Sources coded (%)	Female students Sources coded (%)	Total Sources coded (%)
Reply	11 (36,7%)	19 (63,3 %)	30 (100%)
New entry	7 (36,8%)	12 (63,2%)	19 (100%)

Table 6 shows the students contributions (total and by sex). Not surprisingly, reflection was the most common type of content contribution. It was often the first contribution published by male and female students alike. Affirmation was the second more common type of content contribution, *i.e*., students agreed with the ideas expressed by their colleagues. It was not possible to find significant differences between male and female students in their publishing patterns, considering the relative percentage of male and female students in the course. Content contributions, such as clarification, collegial challenges, personal realizations/transformations, reiteration or substantive insights cannot be taken in consideration because they were less used.

Table 6. Students contributions (total and by sex).

Type of content contribution	Male students Sources coded	Female students Sources coded	Total Sources coded
Affirmation	8	10	18
Clarification	0	3	3
Collegial challenges	2	2	4
Expansive questions	2	4	6
Personal realizations/transformations	1	3	4
Reflection	8	12	20
Reiteration	1	0	1
Substantive insight	3	2	5
Total	25	36	61

5 Discussion

The results obtained by means of a qualitative analysis indicated that the students enrolled in the course had mostly positive attitudes towards the course's online forums (*vide* research question 1). The perceived participation, relevance of the themes, dialogue and meeting others' perspectives, self-expression and improving writing skills were greatly valued.

The content contributions of the students comprehended fundamentally reflections and affirmations, although expansive questions, substantive insights and personal realizations were also found (*vide* research question 2). Currently, another research in which we are involved is trying to understand if it possible to instigate students to employ more complex types of content contributions in order to improve the quality of the forums by asking them to code their own contributions before publishing.

The current research let us confirm that the content contributions categories proposed by Hughes [19] provided a good framework since it was not necessary to add any other category to the list and since all the categories were used during the coding process, excepting the summary/wrap up category (which was intended to capture the teachers' posts) (*vide* research question 3).

We were not capable of finding any significant difference between male and female students' positive attitudes. This is not to say that their evaluation was identical. As a matter of fact, male students were more critical about forums in that they referred as many negative points as their female colleagues.

Overall, forums were probably more interesting to female students (*vide* research question 4). Why is that? The data retrieved from the final synthesis could be biased since it refers to personal views. We had to look at it cautiously, although the qual-

itative analysis suggested that forums provided an environment where female students felt more comfortable to share their thoughts and ideas. This result was in part consistent with Prinsen, Volman and Terwel [22].

The analysis of the most participated forum was important to estimate if female students were in fact more actively engaged in online discussions when compared with their engagement in the practical lessons. We did not expect that they participated more than their male colleagues, but we would hope that they participated at least as actively as them. New entries were as important as threads. Looking at the relative percentage of the female and male new entries and threads we are not able to identify any difference. In fact, the percentage of new entries and threads published by male and female mirrored the percentage of male and female students in the course. This result might mean either that there are no differences between male and female patterns of publishing or that the eventual differences have to be investigated further. The results that we obtained in some content contribution analysis are based upon insufficient cases.

6 Conclusions

In this paper, we were aimed at understanding how Higher Education students perceived their participation on mandatory forums in the context of an optative course and if female students were more actively engaged in the forums.

There is a great dissension on scientific literature on the benefits of digital media and, specifically, on educational settings. Are digital media capable of equalizing and bridging sex-based unbalances?

Our results give us a few reasons to think that forums might be useful to promote personal development in the context of Higher Education and empower female students to actively engage in discussion. In the first place, attitudes towards forums were positive: students considered that they allowed them to meet other's perspectives. This is of utmost importance in a globalized and plural world. In the second place, female participated as much as male in online forums and eventual differences were not identifiable. This is relevant because they did not have a similar behaviour in the practical lessons, but this is also insufficient. Digital media should not only be capable of providing female students with a more comfortable environment but they should empower them to reconfigure their non-virtual participation in discussions. This was, however, beyond the scope of our present research.

This research is considerably limited. The most relevant limitation is that the coding process was done by one analyst alone and, consequently, the validity of the present results was threatened. The fact that we only analysed one single forum, although the most participated one, is another limitation. In fact, one cannot tell for sure that results would be similar to those that we presented in this paper if other forums were analysed.

These are, nonetheless, preliminary results and future developments must necessarily address the limitations that we stressed by asking other analyst to review the

present coding and by including all forums in the analysis, eventually, using the theoretical saturation principle to determine when no further entries need to be analysed. This trend of research also needs to be deepened and, currently, a larger team is trying to understand whether forum contributions are improved by asking students to self-coding their own posts with the categories proposed by Hughes [19] before publishing them and if any sex effect is associated with students performance in these tasks.

Acknowledgements

The author is very grateful to Professor João Paiva and Professor Carla Morais for giving us access to the course; to the students enrolled in the course whose engagement in all the activities was inspiring; to Professor Joan Hughes whose lectures on the 2014 summer course in Lisbon were very insightful; and, finally, to the reviewers, whose comments considerably helped to improve this paper.

References

1. Levin, D., Arafeh, S.: The digital disconnect: the widening gap between Internet-savvy students and their schools. PEW Internet and American Life Project. (2002)
2. European Commission: Survey of schools: ICT in education. Benchmarking access, use and attitudes to technology in Europe's schools (2013)
3. Kolikant, Y. B.-D: Using ICT for school purposes: Is there a student school disconnect? Computers & Education, 59, 907-914 (2012)
4. Chen, W.: A moveable feast: do mobile media technologies mobilize or normalize cultural participation? Seminar. University of Porto, Portugal (December, 15, 2014)
5. DGEEC/MEC, PORDATA. Docentes do ensino superior: total e por sexo – Portugal, http://www.pordata.pt/Portugal/Docentes+do+ensino+superior+total+e+por+sexo-666
6. Sheltzer, J.M, Smith, J.C.: Elite male faculty in the life sciences employ fewer women. Proceedings of the National Academy of Sciences of the United States of America, 111 (28), 10107-10112
7. Reuben, E., Sapienza, P., Zingales, L.: How stereotypes impair women's careers in science . Proceedings of the National Academy of Sciences of the United States of America, 111 (12), 4403-4408
8. Alferes, V. R., Bidarra, M. G., Lopes, C. A., Mónico, L. S.: Domínios de investigação, orientações metodológicas e autores nas revistas portuguesas de psicologia: Tendências de publicação nas últimas quatro décadas do século XX. Análise Psicológica, 27(1), 3-20 (2009)

9. Prinsen, F. R., Volman, M., Terwel, J.: Gender-related differences in computer-mediated communication and computer-supported collaborative learning. 23, 393-409 (2007)
10. Gregory, S.: Discussion boards as collaborative learning tools. International Journal of Continuing Engineering Education and Life-Long Learning 25 (1) (2015)
11. Morais, C., Moreira, L., Paiva, J. C.: Methodological approaches used to study Information and Communication Technologies in Education: a systematic review of the Portuguese scientific production in SCOPUS and Web of Science. ICERI2014 Proceedings, 2176-2183 (2014)
12. Morais, C., Moreira, L., Paiva, J. C.: Methodological approaches used to investigate "online learning". A systematic review of the Portuguese scientific production in SCOPUS and Web of Science. Aprendizagem Online. Atas Digitais do III Congresso Internacional das TIC na Educação, 419-425 (2014)
13. Wang, Q., Woo, H. L., Quek, C. L., Yang, Y., Liu, M.: Using the Facebook group as a learning management system: An exploratory study. British Journal of Educational Technology 43 (3), 428-438 (2012)
14. Jenkins, H.: Confronting the Challenges of Participatory Culture: Media Education for the 21st Century (2006)
15. Thomas, M.J.W.: Learning within incoherent structures: The space of online discussion forums. Journal of Computer Assisted Learning 18(3), 351-366 (2002)
16. Oliveira, I., Tinoca, L., Pereira, A.: Online group work patterns: How to promote a successful collaboration. Computers & Education 57(1): 1348-1357 (2011)
17. Rovai, A. P.: Facilitating online discussions effectively. Internet and Higher Education 10, 77–88 (2007)
18. Schrire, S.: Knowledge building in asynchronous discussion groups: Going beyond quantitative analysis. Computers & Education, 46, 49–70 (2006)
19. Hughes, J.: Types of content contributions. Personal communication (2014)
20. Hsi, S., Hoadley, C. M.: Productive discussion in science: gender equity through electronic discourse. Journal of Science Education and Technology 6 (1), 23-36 (1997)
21. Lim, S. Nahyun, K. Gender differences in information behavior concerning Wikipedia, an unorthodox information source? Library & Information Science Research 32, 212–220 (2010)
22. Prinsen, F. R., Volman, M., Terwel, J.: Effects on participation of an experimental CSCL-programme to support elaboration: Do all students benefit? Computers & Education 52, 113–125 (2009)
23. Bardin, L.: Análise de Conteúdo. Lisboa: Edições 70 (2004)
24. Glaser, B., & Strauss, A. The discovery of grounded theory: Strategies for qualitative research. New York: Aldine De Gruyter (1967)
25. Eagly, A., Chaiken, S.: Attitude structure and function. The Handbook of Social Psychology 1, 269-322. New York: McGraw-Hill (1998)

Context-based learning games for children with cerebral palsy: a prototype

Elis Silva[1], Jorge Silva[2]
[1] Faculty of Engineering of the University of Porto,
University of Coimbra [2]
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal,
{elissilva481@gmail.com, sasilva@dei.uc.pt}

Abstract. Currently, a part of population about one billion, about 15%, including children, are living with a disability, be visual, hearing or physical. People with deficiency can be overly dependent on their families, due the lack of support services. The Internet of things can provide a better life for these people and allows them to participate in the social and economic life. In this paper we describe an application developed for Android, a learning game to be use by children with cerebral palsy in context of learning in the classroom. The work was discussed with a group of specialist of APCC institution, and were identified limitations in the application use and have been proposed new challenges for application. In this paper were also discussed concepts on Wireless Sensor Networks and Internet of Things for people with disabilities

Keywords: Internet of Things, Cerebral Palsy, Wireless Sensor Networks.

1 Introduction

According to the World Health Organization (WHO), a part of population of about one billion, about 15%, including children, are living with a disability, be it visual, hearing or physical [1]. People with disabilities are often dependent on others to carry out everyday activities. The services that enable people with disabilities to have greater autonomy are limited. Therefore, these people cannot be fully integrated into society.

The IoT allows through low-cost applications the democratization and use of these technologies by all social classes. The IoT applications for people with disabilities are interesting, they aim to increase the autonomy of these people, or even to their caregivers [2].

Within this project in IoT, we made a partnership with the Cerebral Palsy Association of Coimbra-APCC. On The needs and areas of intervention collected together the institution, a greater emphasis was given to the shortage of learning games using the scanning method (selection mechanism widely used in games for people with cerebral palsy). Our goal therefore, was the development of a learning game for tablet with Android operating system, where this game involves the interaction of a child using the device with a toy.

In section 2 we describe the concepts of cerebral palsy, IoT and Wireless Sensor Networks. Section 3 describes the implementation of the architecture of Project, and the operation of the application and the technologies used. In Section 4 are presented

the final considerations of APCC institution about the project, and finally, section 5 presents the conclusion and future work.

2 Background

In this section are addressed the main concepts of cerebral palsy, and concepts about Internet of Things (IoT), and in particular applied IoT to deficiency and concepts and work on Wireless Sensor Networks.

2.1 Cerebral Palsy

Cerebral palsy is characterized by a group of disorders that affect the motor system, and the posture control. The Cerebral palsy appears early after an injury, damage or disorder of the central nervous system [3].

The treatment is palliative for cerebral palsy, since it is not possible to act on a healed injury. In addition to drug treatments, the therapies in general are widely used, such as speech therapy. The speech therapy consist the development of activities in scope the prevention, evaluation and treatment of disorders in communication. The Professionals use learning games as assistance in the performance of this therapy. However, there is little availability of these games on the market, the game "The Grid 2" is an example of a game that allows people with limitations in speech can communicate via computer. The Idea Project - Digital Divide with Teaching Interactive Accessible [4], which integrates contents of the 1st Cycle of the areas of Mathematics, Portuguese and Environmental Studies, allows users to use the scanning method for access to games. There are mobile applications, as the example of Proloquo2Go [5], for the iPhone and the Sleep Flex Lite [6] for Android, that use alternative communication as tool to improve the communication capacity of people with disabilities for speak. The applications are similar and allows that is issued a voice with to the selected information.

2.2 Internet of Things

For Domingo [2] the Internet of Things (IoT) is a technological revolution in computing and communications. For [7] the concept of IoT is like a vision where the objects in our world are identified only as part of the Internet, with important information and may be accessed over the network, which has dramatically impact in the professional, personal and social. According to [8], technological change allows a different form of communication between people and the things. Although there are different definitions for the Internet of Things, there are underlying concepts that normally appear when it comes to defining their goals.

The Internet of Things is closer to marking technologies, as the Radio-Frequency Identification (RFID), wireless sensor networks, actuators, mobile phones, the quick response code (QR codes), Near Field Communication (NFC), among others. The interaction and cooperation between these objects will occur only through addressing schemes aiming to achieve common goals, [9]. Among the numerous fields of

application where the technology of the Internet of Things will become very useful, we highlight following scenarios: homes and smart cities. An example of this is the project USEFIL - Unobtrusive Smart Environments for Independent Living [10], which have the goal to generate systems and services through a simplified approach with cost-effective solutions for older people, making use of Information and Communication Technologies (ICT). For smart cities, stands out the European project OpenIOT, that aims to facilitate the use of sensors in ICT-based services not only for smart cities, but also in industry and agriculture with solutions based on sensors networks service [11]. The Commodity project aims to develop an intelligent system for the analysis of combined medical data to make possible the provision of medical information directed to the treatment of a single patient [12].

2.3 Wireless Sensor Networks

Wireless Sensor Networks (WSNs) differ in several aspects of traditional computer networks. Usually consist of many nodes scattered in a region, in order to take measurements of some phenomenon, for example, seismic measurements of a volcano. The collected data are send to a base station, for to be analyzed and treated. The nodes have energy limitations, because in many cases, are in inhospitable regions or of difficult access, and must have mechanisms for self-configuration and adaptation because of losses of nodes (either by destruction of equipment or due to complete loss of their power supply source) or the insertion of new nodes [13].

In WSNs, each node may be equipped with a set of sensors, such as temperature, pressure, humidity, light, sound levels, and other. The combination of these devices provides WSNs be used in a wide range of applications such as [14]: **Agriculture and environment**, where is possible perform the measurement of the fertilizer level in the whole extension of the property. These systems provide to farmer a precision farming leading leading to lower costs, because the farmer just need to perform a new application of fertilizer in deficit regions. **Military applications**, monitoring of strategic regions can be performed by magnetic sensors and vibration thereby allowing identifying moving enemy troops and assist in the decision-making process during the battle. **Medical Applications** enables you to perform the monitoring of vital signs of patients, organ functioning as the heart and detect the presence of substances harmful to man.

2.4 Internet of Things for people with disabilities

The Internet of Things can provide a better life for people, particularly for people who need support, because of their disability. Therefore, the IoT can help these people with special needs to enhance their social life, offering in their daily activities the assistance they need, providing greater autonomy, independence and economic participation. There are many current projects discussing architecture IoT for people with disabilities. In [2] had proposed the IoT architecture for applications aimed at people with disabilities. This architecture was divided into three layers: the first layer

is the perception, the main function is to identify objects and collect sensitive information to the context of the environment of people with disabilities is comprises for sensor means, actuators, tablet, PC, Smartphone, RFID, among others. The second layer is the Network and its main function is the transmission of information obtained from the perception layer. It consists of wireless networks, Internet, and so on. Have the Application layer is a set of intelligent solutions applied to IoT technology, in order to meet the needs of users.

The European Union (EU) in the IOT-i EU project [15] suggests another proposal of IoT architecture. This project aims to create a unified system for IoT community, aiming to align the community members in a vision of IoT exclusively for the Internet of the future. Thus, it was intended to avoid causing the fragmentation the IoT in several different solutions of the application domain. However, one of the first projects worldwide to IoT was the SENSEI FP7, whose goal was to build an architecture that allows the integration of sensor networks and wireless actuators. The work influenced other projects about architecture of the Internet of Things, such as the IOT-i EU FP7 previously described, and FI-WARE project with the main objective of creating APIs (Application Programming Interface) open to developers and suppliers through a common architecture.

3 System Overview and Architecture

In this section we will present a overview of the application and the architecture the project developed, as well as the technologies used.

3.1 Application and General Architecture

The android application is a learning game called "Game of Animals" designed for use by children with cerebral palsy. Designed for tablets with android operating system. In this application were made initially the settings by tutor. This setting will establish communication with the multithreaded server and from that will define the set of questions that will appear in the game to the student. With these initial settings completed, students can learn about a number of animals previously defined in the application, and when deemed necessary start the exercises through a Quiz system.

To carry out the management of the activities with the students, the tutor (teacher / parent) has on your domain a system called Tutor System. This system is capable of perform administering all the necessary resources so that each student can receive adequate exercise to their learning needs.

In Figure 1 is show the project overall architecture. This architecture consists of the android and tutor applications, by multithreaded servers and UDP, besides the sensor nodes. In the following paragraphs are briefly described each module of architecture.

The multithreaded server is responsible for providing all questions and alternatives in carrying out activities in the Animal Game. It also gets all the answers that the student chose during the course of exercise, inserts the data in the database and finally sends the UDP server what action should be performed by the toy node. Upon receiving the request from the multithreaded server, the UDP server is responsible

only for forwarding this request by port serial device to the base station, which in turn forwards via radio to toy node, which should be your action according to the response of each student. Finally, upon receiving the message, the toy node performs the action of turn on the green light when the answer is right and turn on the red light when the answer is wrong.

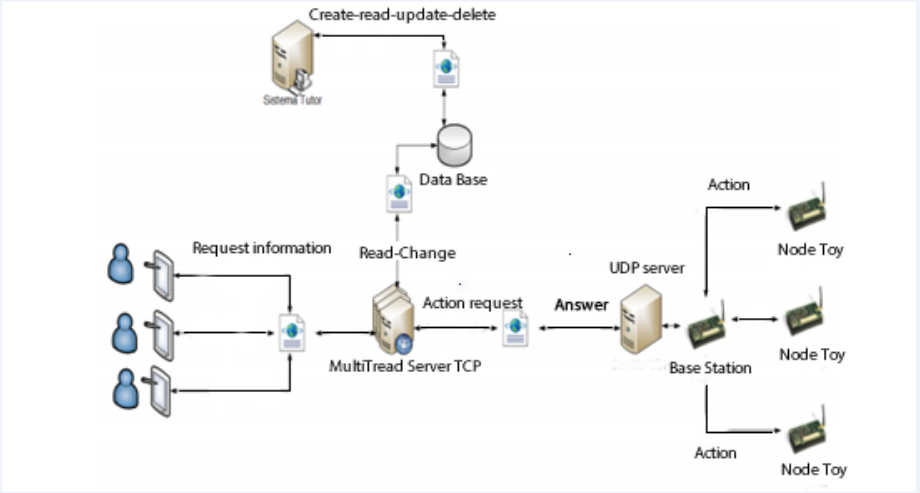


Figure 1 - General architecture of the Application

3.2 Technologies Used

We have used Java programming language for development the android application. We use sensors MicaZ (Datasheet MicaZ), which are third-generation devices that enable low power consumption, and are able to work with wireless sensor networks in 2, 4 GHz e 868/916 MHz [16]. In addition, support the date transmission rate of 250 kbps. These nodes also have temperature sensors, humidity, and pressure, among others. We also used a MicaZ base station that was connected to the Linux server (Ubuntu operating system), and so was able to receive via serial port application android information and send action commands to the sensor node.
In the project was used the MySQL 5.5.29 database to store game information. The management program date base was choosing was due mainly to be open source.The choosing this management program was due mainly to be open source, besides having excellent performance and stability, requires little hardware resources [17].

3.3 Multithreaded Server

To make it possible verify the actions and intentions that the student or tutor wanted done while executing the application android.was created a communication client-server through socket able to meet all the needs of users for example: verify that the answers are correct, provide the questions customized to the student that is playing and save the student's performance in the database. Also, being able to provide at the same time this service to more than one user. The message that starts the operation, part of the android application system when it starts to perform the setup procedure the game environment by tutor. The message sent is responsible for establishing communication via socket with the multithreaded server. Then the tutor forwards the code identification of the student to the server. On receiving the message, the server checks whether the ID code of the student is valid, then forwards for android application all groups in which the student is associated.

The tutor makes the pick of the bunch, and mainly the type of activity that the tutor want to accomplish with the student can motivate this choice. Upon receiving the chosen group, the server forwards all the issues and alternatives associated with the group for the android application. Having all the information necessary for the completion of the quiz is then available the opportunity to start the game. For each question in the quiz, the android application forwards the response to the server.

Then the server identifies whether the answer is correct or incorrect and make available for android application the updated score of the student, along with the name and at most the score of 2 students who belong to the same group. This repeats until the last question is answered.

3.4 UDP server

The UDP server the same way that the multithreaded TCP server also is implemented via socket, using mostly different communication protocols. The choice of the UDP protocol instead of the TCP protocol, to performing communication between two servers, is justified by the fact that protocol UDP is shown more appropriate for data stream in real-time, thereby being more suiting to ensure the concept of cause and desired effect in the system.

The message that starts communication between the two servers, is sent to UDP server after multithreaded server check if the response sent by android application is correct or wrong. To perform the communication has been established two types of messages, one for when the response is correct, being asked that the green light turn on by the toy node, and another message when the answer is wrong is requested that the red light is lit. After receiving one of two messages, The UDP server encapsulates the message through their own TinyOS library that provides a Java interface at the application level for sending messages, and forwards, via serial port to the base station.

3.5 MicaZ

Described above, for the construction of the infrastructure of hardware necessary for perform the interaction among the application and the toy, we used two MicaZ devices for construction of this architecture. A device is used to serve as a base station and assist in communication and sending messages to other software. The second device was meant to serve as own toy node and create interface with the student during the game. The activity begins after the base station via the serial port receive one message from the UDP server, then begin the message reading process, being initially identified for which toy node this message refers. From this, each student can interact with your toy of shape independent of the others. Before starting any message forwarding process for toy, a check is performed if can use the of system radio of device, otherwise, this request is then routed to a queue until their use is allowed. With the availability of all resources necessary to initiate communication with the toy, is then created a message from the below written communication structure. In this structure two attributes bases are defined, one for the identification of the toy node and the second for the message content.

Table 1 - Communication Structure

```
typedef nx_struct BlinkToRadioMsg {
    nx_uint16_t nodeid;
    nx_uint16_t counter;
} BlinkToRadioMsg;
```

For the filling up of message content is checked what action was originally requested by the UDP server, and then represent it in a 16-bit integer, being assigned a value of "0x01" to turn on the red light and the value of "0x02" to turn on the green light. With all the completed fields is carried the send of message for the toy.

4 Considerations APCC

In the final phase of the project, was necessary to perform a presentation to the APCC professionals of, for to understand the functioning of the developed system. Were had presented the tutor system and the android application, showing systematically the features of each application, such as example of a registration of a new student in the Tutor application as also the action triggered for toy node to each right or wrong in the Animal Games.

Were presented some limitations in the application. These limitations refer to some requirements that were not possible to develop in this project, as an example: the application to be tested using a switch to trigger some information of the game. This occurred, because we do not had that equipment for the tests. Another limitation that

do not posed was the toy with sound feedback and move. This toy would use to promote a stimulus reinforcement in the game with the children. The application developed for the institution presents the needs raised in the beginning of this project. These needs are mainly related to the motor difficulties of children with cerebral palsy to access applications on your computer, requiring mechanisms that allow such access, as the use of switches and adapted keyboards.

Was very important to development tutor system to use by professionals who are to accompany the student during an activity. In APCC is currently not used an application that allows this type of management, because through this application, the tutor can perform several actions to make custom game for each child's learning needs.

APCC are used various games, such as The Grid 2 that is used in therapies talking and activities in the classroom with the using the switch the use of switch simplifies much the lives of people with motors commitments, for this reason are widely used in various activities of daily life, the example of the wheelchair handling.

Just as is done training with children to use the switch, is being developed recently training in the institution for the use of the Tablet. This training is to establish small advances with handling hand to touch the device screen. Yet is not used any application developed for Tablet that has the scanning mechanism, and which allows the use of the switch, because the institution is unaware of any such application. The animal game developed in android is a good starting point for the creation of several other applications that can be used by children with cerebral palsy.

In general, the game has several features not found in other; one example is the group games. The functionality of the group of play is not intend to encourage any competition among students, because each has its difficulties and limitations, but allow greater socialization of students in the classroom.

Another feature of the application is the inclusion of actual toys that interact with the application in android. These light stimuli fired every right or wrong during the game arouse students' attention and interest in learning the game. The prototype presented of the application with the interaction with the toy not properly done; it lacked the coupling of the sensor nodes that trigger the light effects in the toy. It is expected that in addition to the light effects are also presented sound effects and movement in the toy every learner response.

Although some limitations that were not implemented in the project, the implementation involves a large part of the requirements raised by the institution to develop an application that can be used by students in the APCC.

5 Conclusion and Future work

In scope of the Internet of Things in the development of learning games with toys coupled with sensors nodes, an application was created in android. This game offers a simple interface with mechanisms that allow children to have access to same. The access mentioned here refers to mechanism sweep just mentioned. The application was developed in accordance with the set of requirements drawn up together with APCC institution. We show here in the project not only the application for children, but also a dedicated application for tutor. The application prototype has been tested and shown to the professionals of the institution, and so allowed us to conclude that the system developed for the APCC meet most of the requirements mentioned.

The developed application differs from the games already used by the institution, at first was developed specifically for devices, such as tablets and smartphone, that for the moment are not used by APCC any game available for use in these devices. The advantage of an application used in Tablets are related to portability of the device, that is, allow the same may be coupled in a wheelchair for children, and being smaller than a laptop and lighter, this factor has become paramount in the implementation of usability.

In this work the main objective was fulfilled, to create an application at low cost to allow monitoring by tutors or parents in the development of the activities carried out by children at play performed in the classroom and also at home, and even the inclusion of devices represented here by us toys in the application, interacting in real time with the virtual game, according to the answers given by the students. The IoT in this context was important because it allowed the integration of the actual represented by toy with sensor together with the application in android.

For possible future work, we highlight the creation of new functionality of interaction with the toy. These features relate particularly in developing more strengthening mechanisms of stimuli such as sounds and movements. The tutor application will be important to develop new features to make it more manageable application by tutors. An important issue is the possibility of the tutors can create custom questions may choose images and sounds that appear in the Animal Game.

In this project was not possible to use a switch to test the operation of the android application, we used the touch on the device screen to trigger the information. Will be very important for the next work the switches to be incorporated (switch) as triggering devices for children, as it will facilitate their access in the use of applications in android.

References

1. W. H. Organization. [Online]. Available: http://www.who.int/en/. [Accessed in 02 10 2013].
2. M. C. Domingo, "An overview of the Internet of Things for people," Journal of Network and Computer Applications , vol. 35, pp. 584-596, 2012.
3. G. P. Fernandes e J. M. S. L. Resende, "Paralisia cerebral Aspectos Fisioterapêuticos e Clínicos," Neurociências, vol. 12, n.º 1, p. 41, 2004.
4. iD032, "IDEIA - Inclusão Digital com Ensino Inte-rativo," [Online]. Available: http://www.acessibilidade.gov.pt/conferencia_id/id032.html. [Accessed in 20 10 2013].
5. AssistiveWare, "AssistiveWare," A voice for those who cannot speak: Proloquo2Go, [Online]. Available: http://www.assistiveware.com/product/proloquo2go. [Accessed in 21 11 2013].
6. G. Play, "Sono Flex Lite," Sono Flex Lite,[Online]. Available: https://play.google.com/store/apps/details?id=com.tobii.sonoflexlite&hl=pt_PT. [Accessed in 22 10 2013].
7. J. Eksteen e L. Coetzee, "The Internet of Things-Promise for the Future," In Proceeding of the IST-Africa Conference, pp. 1-9, 11 5 2011.
8. W. N. Tan L, "Future internet: the internet of things," em international conference on advanced computer theory and engineering (ICACTE'10), Chengdu, , 2010.
9. J. R. Molina, J. -F. Martínez, P. Castillejo e L. López, "Combining Wireless Sensor Networks and Semantic Middleware for an Internet of Things-Based Sportsman/Woman Monitoring Application," Sensores, vol. 13, pp. 1787-1835, 2013.
10. "USEFIL," USEFIL, [Online]. Available: http://www.usefil.eu/. [Accessed in 02 November 2013].
11. "OPENIoT," Open Source Solution for the Internet of Things into the Cloud, [Online]. Available: http://openiot.eu. [Accessed in 22 11 2013]..
12. COMMODITY, "COMMODITY," Welcome to the Commodity12 Project-COntinuous Multi-parametric and Multi-layered analysis Of Diabetes Type 1 & 2, [Online]. Available: www.commodity12.eu/. [Accessed in 22 11 2013].
13. A. A. Loureiro, J. M. S. Nogueira, L. B. Ruiz, R. A. d. F. Mini, E. F. Nakamura e C. M. S. Figueiredo, "Redes de Sensores Sem Fio," XXI Simpósio Brasileiro de Redes de Computadores, pp. 179-226, 10 12 2003.
14. I. M. Mohammad Ilyas, Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press;, 2004.
15. IoT initiative (IoT-i), [Online]. Available: http://www.iot-i.eu/. [Accessed in 11 11 2013].
16. memsic, [Online]. Available: http://www.memsic.com.php5-12.dfw1-1.websitetestlink.com/products/wireless-sensor-networks/wireless-odules.html. [Accessed in 10 01 2014].
17. Mysql, 20 01 2013. [Online]. Available: http://www.mysql.com/. [Accessed in 20 01 2013].

SESSION 2

KNOWLEDGE DISCOVERY

A Survey of Merging Decision Trees Data Mining Approaches
Pedro Strecht

A Review of recent progress in multi document summarization
Shazia Tabassum and Eugenio Oliveira

Text Mining Scientific Articles using the R Language
Carlos Gulo and Thiago Reis

Characterizing Developers' Rework on GitHub Open Source Projects
Thiago Reis and Carlos Gulo

A Survey of Merging Decision Trees Data Mining Approaches

Pedro Strecht

INESC TEC/Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
pstrecht@fe.up.pt

Abstract. The merging of decision tree models is a topic lacking a general data mining approach that is not domain specific. Existing research address the issue under different motivations and to solve different problems. This paper presents a survey of current approaches of merging decision trees, highlighting what they share in common by presenting a general data mining approach based of the combination of rules. Although its major components and problems are abstracted, illustrative examples from the literature are provided. Possible directions of unexplored issues for future research are also discussed.

Keywords: prediction models, decision tree merging, survey

1 Introduction

Classifiers obtained from decision tree models have the characteristic of not requiring previous domain knowledge or heavy parameter tuning making them appropriate not only for prediction but also for exploratory data analysis. The tree-like representation of knowledge presents itself as intuitive, making models that are usually interpretable by humans [1]. For this reason, decision trees models have been very popular as models in classification problems in various business domains and are still widely used.

The motivation to merge models has its origins as a strategy to deal with building prediction models for distributed data. Distribution can occur naturally, i.e., when the data is initially collected on distributed locations and transportation to form a *monolithic data set* (designation used in literature to refer to a single centralized data set) is costly or unsafe making it not feasible. An example is Bursteinas and Long [2] motivation which is related to data being generated on distributed distant machines connected by low transparency connections. Alternatively, distribution can occur artificially, being a strategy to deal with very large monolithic data sets which would make training a model a very slow task or even impossible due to lack of resources. Data sets exceeding RAM sizes is presented as a factor for distributed data by Andrzejak, Langner and Zabala [3]. Another reason to have artificially distributed data is when it is collected as consequence of a business process. An example is in Strecht, Moreira and Soares [4] research in which student enrolments records are gathered in courses

to create models at course level in a university. Merging models appears as a technique to generalize the knowledge contained in those models at university level to provide useful information to help explain the drop out phenomenon.

A *local data set* can be defined as a subset of a larger monolithic data set that is splitted either naturally or artificially. Each local data set provides training examples to create local models. If the number of models is too large, it becomes difficult to generalize knowledge, and have a single model view. There are two major approaches to build generalized models from distributed data:

- *Data compression* in which data in each local data set is compressed into one or more training examples. Another process then gathers all examples and trains a generalized model. Yael and Elad [5] describe this approach in detail. The drawback is that there are no models created at each local data set, which may be required to understand business processes at local level.
- *Model merging* in which a local model is trained in each local data set and then another process combines them to form a generalized model. This has been used to a greater extent with different approaches which can be divided into two main groups: mathematical, in which a mathematical function is used to aggregate models; and data mining in which the models are broken down into parts, combined and re-assembled to form a new model (both are detailed in Section 2).

At first glance, it may seem that merging models is another form of ensemble learning. There are, however, major differences between the two methodologies that help to easily distinguish them, as presented in Fig. 1.

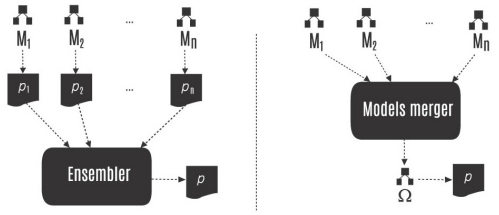


Fig. 1. Ensemble learning vs Models merging

Ensemble learning consists in combining the predictions (p_1, ..., p_n) made by various models (M_1, ..., M_n) into one prediction (p). The ensembler implements a method to combine the predictions, such as bootstrap aggregating (bagging), boosting, or random forests. Models merging consists in combining various models (M_1, ..., M_n) to create a merged model (Ω) which is the only one making a prediction.

The remainder of this paper is structured as follows. Section 2 presents an overview of approaches to merge decision trees. Section 3 describes a general data mining approach with different alternatives of carrying out intermediate steps. Section 4 presents the conclusions and identifies open issues for research.

2 Overview of Approaches for Merging Decision Trees

2.1 Mathematical Approaches

Kargupta and Park [6], motivated by the need to analyse and monitor time-critical data streams using mobile devices, propose an approach to merge decision trees using the Fourier Transform. This mathematical operation decomposes a function of time (a signal) into its frequencies yielding the frequency domain representation of the original signal. According to the authors, mining critical data streams requires on-line learning that produces a series of decision trees models, which may have to be compared with each other and aggregated, if necessary. Transmitting these models over a wireless network is presented as a problem. As the decision tree is a function, it can be represented in frequency domain, resulting in the model *spectra*. Merging models becomes a matter of adding theirs spectra's, a trivial task in frequency domain. If required, the merged model can be transformed back to the decision tree domain by the Inverse Fourier Transform. This approach has not evolved since 2004, when it was first presented. It has been criticized [3] for being difficult to extend (e.g. only binary attributes are considered by the authors) and for the lack of performance measures.

Gorbunov and Lyubetsky [7] address the problem by proposing a mathematical approach to construct a decision tree that is closest on average to a set of trees. The problem is analysed from a theoretical point of view, i.e., it is not presented as a solution to be used in a specific application. Nevertheless, decision trees illustrating the theory of evolution are pointed out as a proof of concept. A complex algorithm is described and exemplified by a case in which ten binary decision trees are combined, resulting in a *super tree* that represents their average. The algorithm is cited afterwards by the authors in another research [8] in the context of molecular biology, therefore suggesting what seems to have been the main motivation for its development. Although developed quite recently, in 2011, it has not been used by other researchers, probably due to its complexity.

2.2 Data Mining Approaches

Provost and Hennessy [9, 10] present an approach to learning and combining rules on disjoint subsets of a full training data. A rule based learning algorithm is used to generate rules on each subset of the training data. The merged model is constructed from satisfactory rules, i.e., rules that are generic enough to be evaluated in the other models. All rules that are considered satisfactory on the full data set are retained as they constitute a superset of the rules generated when learning is done on the full training set. This approach has not been replicated by other researchers.

A more common approach is the combination of rules derived from decision trees. The idea is to convert decision trees from two models into decision rules by combining the rules into new rules, reducing their number and finally growing a decision tree of the merged model. The basic fundamentals of the process are first presented in the doctoral thesis of Williams [11] and over the years, other

researchers have contributed by proposing different ways of carrying out intermediate tasks. Table 1 summarizes research examples of this approach, specifying the problem (or motivation) and data sets used.

Table 1. Research examples of combination of rules approaches to merge models

Research	Problem/motivation	Data sets
Hall, Chawla and Bowyer [12]	Train model in a very large data set	Iris, Pima Indians Diabetes
Bursteinas and Long [2]	Mining data distributed on distant machines	UCI Machine Learning Repository
Andrzejak, Langner and Zabala [3]	Train models for distributed data sets and exceeding RAM sizes	UCI Machine Learning Repository
Strecht, Moreira and Soares [4]	Generalize knowledge in course models at university level	Academic data from University of Porto

Hall, Chawla and Bowyer [12, 13] research present as rationale that is not possible do train decision trees in very large data sets because it could overwhelm the computer system's memory by making the learning process very slow. Although a tangible problem in 1998, nowadays, this argument still makes sense as the notion of very large data sets has turned into the big data paradigm. The approach involves breaking down a large data set into n disjoint partitions, then, in parallel, train a decision tree on each. Each model, in this perspective, is considered an independent learner. Globally, models can be viewed as agents learning a little about a domain with the knowledge of each agent to be combined into one knowledge base. Simple experiments to test the feasibility of this approach were done on two datasets: Iris and Pima Indians Diabetes. In both cases, the data sets were split across two processors and then the resulting models merged.

Bursteinas and Long [2] research aims to develop a technique for mining data which is distributed on distant machines, connected by low transparency connections arguing that there is a lack of algorithms and systems which could perform data mining under such conditions. The merging procedure is divided into two scenarios: one for disjoined partitions and one for overlapped partitions. To evaluate the quality of the method, several experiments have been performed. The results showed the equivalence of combined classifiers with the classifier induced on a monolithic data set. The main advantage of the proposed method is its ability to induce globally applicable classifiers from distributed data without costly data transportation. It can also be applied to parallelise mining of large-scale monolithic data sets. Experiments are performed merging two models in data sets taken from the UCI Machine Learning Repository [14].

Andrzejak, Langner and Zabala [3] propose a method for learning in parallel or from distributed data. Factors cited as contributing to this trend include emergence of data sets with exceeding RAM sizes and inherently distributed scenarios such as mobile environments. Also in these cases interpretable models are favoured: they facilitate identifying artefacts and understanding the impact of individual variables. The method is compared with ensemble learning, because

in a distributed environment, even if the individual learner on each site is interpretable, the overall model usually is not, citing as example the case of voting schemes. To overcome the problem they propose an approach for merging of decision trees (each learned independently) into a single decision tree. The method complements the existing parallel decision trees algorithms by providing interpretable intermediate models and tolerating constraints on bandwidth and RAM size. The latter properties are achieved by trading RAM and communication constraints for accuracy. The method and the mentioned trade-offs are evaluated in experiments on data sets from the UCI Machine Learning Repository [14].

In all previous presented research examples, decision trees were trained using C4.5 algorithm [15] and accuracy [1] was used as evaluation function of the individual and merged models.

Strecht, Moreira and Soares [4] research on educational data mining starts from the premise that predicting the failure of students in university courses can provide useful information for course and programme managers as well as to explain the drop out phenomenon. The rationale is that while it is important to have models at course level, their number makes it hard to extract knowledge that can be useful at the university level. Therefore, to support decision making at this level, it is important to generalize the knowledge contained in those models. An approach is presented to group and merge interpretable models in order to replace them with more general ones without compromising the quality of predictive performance. The case study is data from the University of Porto, Portugal, which is used for evaluation. The aggregation method consists mainly of intersecting the decision rules of pairs of models of a group recursively, i.e., by adding models along the merging process to previously merged ones. The results obtained are promising, although they suggest alternative approaches to the problem. Decision trees were trained using C5.0 algorithm [16] and F1 [17] was used as evaluation function of the individual and merged models.

3 Combination of Rules Approach to Merge Models

The *combination of rules* approach to merge decision trees models is the most common found in the literature. However, when presented it has always been in the context of a specific problem intertwined with details from the context of business rules. Therefore, there is the lack of a generalized approach that is not restricted to a specific domain. This section proposes such an approach by identifying its key components and the major problems encountered. Aligned with a survey perspective, it also presents the different alternatives to carry out intermediate tasks found in the literature.

Fig. 2 presents the system architecture of this approach which encompasses four main processes: models creation and evaluation, models grouping, models merging and group models evaluation. The local data sets (D_1, ..., D_n) are assumed to be collected and prepared by a data extraction process which is not part of the methodology. The outputs are group models (G_1, ..., G_k) and corresponding performance measures ($\theta(G_1)$, ..., $\theta(G_k)$). θ denotes a function

for evaluation measure (e.g. accuracy or F1). The following sub-sections detail each of the processes.

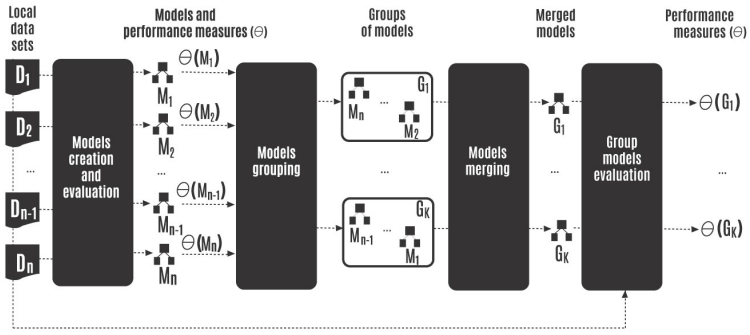


Fig. 2. System architecture of the combination of rules approach to merge models

3.1 Models creation and evaluation

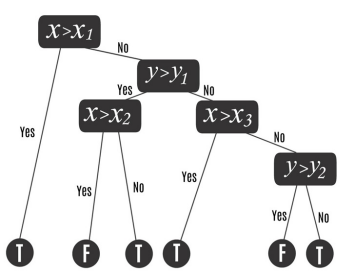
In the first process a decision tree model is created for each local data set. Due to limitations of space, the description of the decision tree induction process is not included. There are several algorithms to create decision trees, the most popular being CART [18] (Classification and Regression Trees) and C5.0 [16] (an evolution of the C4.5 [15] which is an extension of ID3). Although the approach for merging decision trees is not specific to any algorithm, it is recommended that the same algorithm is used to train all individual models. As pre-requisite for merging, it is mandatory to have access to the models themselves. Decision tree algorithms may output the model graphically, through oriented graphs, or textually, as a set of lines. Fig. 3 shows an example of a model in both representations with variables x and y, and classes T and F.

It is worthwhile noting that there may be local data sets for which a model is not created (because of the characteristics within the data). As a consequence, there may be less models than local data sets. For simplicity, however, in Fig. 2 it is assumed that to all n local data sets (D_1, ..., D_n) there is a respective decision tree model (M_1, ..., M_n).

The performance of each model is determined by an evaluation function $\theta(M_i)$. All models should be evaluated using the same experimental set-up and evaluation function.

3.2 Models grouping

In the second process the models are gathered into groups. Although this can be done according to any criteria it is important to make the distinction between two cases of grouping:




```
x > x1: T
x <= x1
:... y > y1
     :... x > x2: F
          x <= x2: T
     y <= y1
     :... x > x3: T
          x <= x3
          :... y > y2: F
               y <= y2: T
```

Fig. 3. Graphical and textual representation of a decision tree model

- *domain-knowledge* in which models are grouped together according to meta-information of the local data sets (e.g. different locations belonging to the same geographical area);
- *data-driven* in which models are grouped together according the characteristics of the models themselves (e.g. measures over the variables used).

Hall, Chawla and Bowyer [12] and Bursteinas and Long [2] do not address the issue of grouping which suggests that the models are all merged in sequence. Andrzejak, Langner and Zabala [3] define a k parameter to study the impact of increasing the number of groups. Each group always has the same number of models, with $k = 1$ being the baseline case. Strecht, Moreira and Soares [4] perform experiments grouping models relating to courses by ten scientific areas (domain-knowledge grouping), by the number of variables and importance of variable (data-driven grouping). The latter by clustering models (with k-means algorithm) using the C5.0 algorithm measure of importance of variable I_v which relates to the percentage of examples tested in a node by that variable in relation to all examples. Finally, a baseline case (similar to the other researchers) was included by forming only one group containing all models.

In Fig. 2, for illustrative purposes, M_n and M_2 belong to the first group (G_1) while M_{n-1} and M_1 are placed in the last (G_k) by some arbitrary criterion.

3.3 Models merging

In the third process the models in each group (or all in sequence if no grouping is performed) are merged together yielding the *group model*, according to the experimental set-up presented in Fig. 4.

A requirement for this process is that each model must be represented as a set of decision rules. This takes the form of a decision table, in which each row is a decision rule. Therefore, the first (M_1) and second (M_2) models are converted to decision tables and merged, yielding the Ω_1 model, also in decision table form. Then the third model (M_3) is also converted to a decision table and is merged with Ω_1 model yielding the Ω_2 model. This process is replicated to all models in the group. The last merged model Ω_{n-1} is converted to decision tree (renamed as group model). Each one of these sub-processes and its tasks are detailed next.

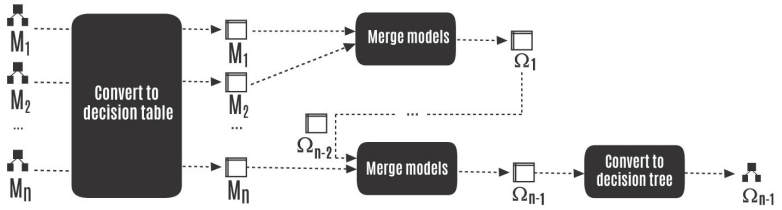
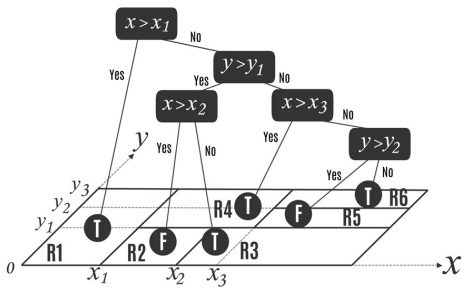


Fig. 4. Experimental set-up to merge all models in a group

Conversion to Decision Table. In the first sub-process, a decision tree is transformed to a set of rules. Each path from the root to the leaves creates a rule with a set of possible values for variables and a class. These have been called "rules set" by Hall, Chawla and Bowyer [12], "hypercubes" by Bursteinas and Long [2], "sets of iso-parallel boxes" by Andrzejak, Langner and Zabala [3], or "decision regions" by Strecht, Moreira and Soares [4]. All these designations arise from the fact that a variable can be considered as a dimension axis in a multi-dimensional space. The set of values (nominal or numerical) is the domain for each dimension and each rule defines a region. It is worth noting that all rules lead to regions that do not overlap and together cover the entire multi-dimensional space.

A *decision table* is the linearisation of a decision tree, being an alternative way of representing it. Fig. 5, in the left, extends the example of Fig. 3 which now includes a two-dimensional space and the corresponding projections as decision regions ($R1$, ..., $R6$). In the right, these are listed in a decision table with columns specifying the class assigned to each region and the set of values of each variable.



Region	Class	Variables	
		X	Y
R1	T	$[0, x_1)$	$[0, y_3]$
R2	F	$[x_1, x_2)$	$[0, y_1)$
R3	T	$[x_2, x_4]$	$[0, y_1)$
R4	T	$[x_1, x_3)$	$[y_1, y_3]$
R5	F	$[x_3, x_4]$	$[y_1, y_2]$
R6	T	$[x_3, x_4]$	$[y_2, y_3]$

Fig. 5. Converting a Decision Tree to Decision Table

Strecht, Moreira and Soares [4] included an extra column referring to the *weight* of a region, which specifies the proportion of examples used by C5.0 to create a region relative to the local data set. In the implementation of the

algorithm that was used, this information is included in each branch of the textual representation of decision trees. The weight is, therefore, a measure of region importance. It is used during the merging process as a strategy to avoid very complex merged models. This is done by only keeping in the merged model the most important regions, i.e., the ones created by a larger number of examples in the original models.

Merge Models. In the second sub-process, two models are merged together which encompasses three sequential tasks. These have been given different designations in the literature. Fig. 6 presents them using Strecht, Moreira and Soares [4] terminology of intersection, filtering and reduction, described next.

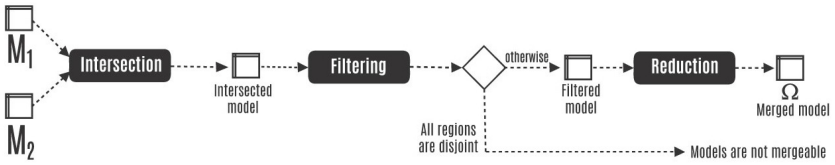


Fig. 6. Sub-process of merging two models

Intersection is a task to combine the regions of two models using a specific method to extract the common components of both, presented in decision table form. The output is the intersected model also in decision table form. The set of values of each region on each model are compared to discover common sets of values across each variable (mapped as a dimension in a multi-dimensional space). The intersection of values of each variable from each pair of regions may have the following outcomes:

- If there are common values, then these are assigned to that variable in the merged region;
- If there are no common values, then they are considered *disjoint regions*, regardless of other variables in which the intersection set may not be empty.

In Hall, Chawla and Bowyer [12] and Strecht, Moreira and Soares [4] approaches, all regions of both models are compared with each other. Bursteinas and Long [2] have a similar method but separate disjoint from overlapping regions. Andrzejak, Langner and Zabala [3] call this operation "unification" and propose a line sweep algorithm to avoid comparing every region of each model. It is commonly based on sorting the limits of each region and then analysing where merging can be done. However, this method only applies to numerical variables.

The class to assign to the merged region is straightforward if the pair of regions have the same class, otherwise the *class conflict problem* arises. Andrzejak, Langner and Zabala [3] propose three strategies to address this problem. The first assigns the class with the greatest confidence, the second, the one with the

greater probability and a third strategy, which is the more complex, involves more passes over the data. Hall, Chawla and Bowyer [12] explore the issue in greater detail and propose further strategies, e.g., comparing distances to the boundaries of the variables. However, this approach seems suitable only for numerical variables. Bursteinas and Long [2] use a different strategy by retraining the model with examples for the conflicting class region. If no conflict arises, that class is assigned, otherwise the region is removed from the merged model. Strecht, Moreira and Soares [4] use the weight associated with each region to decide which is the class to be assigned to the merged region. They study both the impact of choosing the class of the region that has the maximum weight and the minimum weight.

Filtering is a task to remove disjoint regions from the intersected merged model yielding the filtered merged model. Andrzejak, Langner and Zabala [3] call this operation "pruning" and developed a ranking system retaining only the regions with the highest relative volume and number of training examples. Hall, Chawla and Bowyer [12] only carry out this phase to eliminate redundant rules created during the removal of class conflicts. Bursteinas and Long [2] mention the phase but do not provide details on how it is performed. Strecht, Moreira and Soares [4] addresses the issue by removing disjoint regions, recalculating the weight of the remaining ones and highlighting the cases where models are *not mergeable* if all regions are disjoint.

Reduction is a task to limit the number of regions in the filtered merged model, to obtain a simpler model. The regions are examined to find out which can be joined into one. This is possible when a set of regions have the same class and all variables have equal values except for one. In the case of nominal variables, reduction consists on the union of values of that variable from all regions. In the case of numerical variables, reduction is performed if the intervals are contiguous. Another consequence of the reduction is that there may exist variables with the same value in all decision regions. The columns for these variables are removed from the table. Both Strecht, Moreira and Soares [4] and Andrzejak, Langner and Zabala [3] perform this operation. It is not mentioned in the researches of Bursteinas and Long [2] and Hall, Chawla and Bowyer [12].

Conversion to decision tree. In the third sub-process, the last merged model of the group (Ω_{n-1}), in decision table form, is converted to the decision tree representation. Andrzejak, Langner and Zabala [3] attempt to mimic the C4.5 algorithm using the values in the regions as examples. One problem with this method is that it is necessary to divide one region in two to perform the splitting, which increases their number, thus making the model more complex. Hall, Chawla and Bowyer [12] do not perform this phase and the merged model is represented as the set of regions. Bursteinas and Long [2] claim to grow a tree but do not describe the method. Strecht, Moreira and Soares [4] grow a tree of the merged model by generating examples provided from each of the decision regions and submitting them as learning examples to the same algorithm used to create the initial models (taking into account the region weights if desired).

3.4 Group models evaluation results

In the fourth process the group models are evaluated using the same evaluation function used to evaluate the original models. Each research evaluates the group models differently. Hall, Chawla and Bowyer [12, 13] compare the accuracy of the merged model with a baseline model trained with all examples. They observed a slight improvement (about 1%) by using the merged model. Andrzejak, Langner and Zabala [3] also use the same baseline case and then compare its accuracy on increasing the number of groups. They observe that creating up to sixteen groups is the limit where the quality of predictions of the merged model still provides a good approximation to the baseline case. Bursteinas and Long [2] compare the classification accuracy of the test set for the combined tree claiming it to be similar to the accuracy generated with the tree induced on the monolithic data set. Strecht, Moreira and Soares [4] method for evaluation is the most different in the literature. As $F1$ was used as evaluation function in each model, $\Delta F1$ is defined as the gain in the predictive performance by using the group model instead of the original model in relation to each local data set. Also, a *merging score* is defined as the number of models that is possible to merge divided by the number of models in a group. They observed that, merging groups across scientific areas yields, in average, an improvement of 3% in prediction quality.

4 Conclusions

The approach of merging models by combining decision rules is the most often found in the literature. However, being specific to each research, it still lacked a general domain-free specification as the one presented in this paper. There is also the lack of a general terminology for concepts that this paper can partially fulfill.

Grouping models is open for further exploration as it is only tackled by two researches. An important issue that is yet to be investigated is the merge order of models within a group. The merging operation is not commutative, therefore it is of great interest that this aspect should be studied in future research as well as its impact in improving the predictive quality. The process of merging models in a group presents great variations in how it is implemented in the literature. While all agree that the most suitable representation to merge models is working with decision tables (although under different designations), the combination of decision rules algorithm (the core of the whole process) is where the major differences are found. The main problem to deal with is class conflict in overlapping rules that has no consensual approach. Efforts to simplify the resulting merged model are always included mainly by attempting to reduce the number of decision rules. The final sub-process of growing a decision tree representation of the merged model also presents challenges and should be further explored in future research. There is still no consistent way to assess the quality of the merged models or which are the best evaluation measures and experimental set-ups. It is also notable the absence of a common baseline case to be used in any study and help comparison of results.

Finally, existing research shows that merging decision trees, despite being an emerging topic, offers interesting prospects which can make it an exciting research area to be further explored on theoretic and application perspectives.

References

1. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann, 2011.
2. B. Bursteinas and J. Long, "Merging distributed classifiers," in *5th World Multiconference on Systemics, Cybernetics and Informatics*, 2001.
3. A. Andrzejak, F. Langner, and S. Zabala, "Interpretable models from distributed data via merging of decision trees," *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Apr. 2013.
4. P. Strecht, J. Mendes-Moreira, and C. Soares, "Merging Decision Trees: a case study in predicting student performance," in *Proceedings of 10th International Conference on Advanced Data Mining and Applications*, pp. 535–548, 2014.
5. B.-H. Yael and T.-T. Elad, "A Streaming Parallel Decision Tree Algorithm," *Journal of Machine Learning Research*, vol. 11, pp. 849–872, 2010.
6. H. Kargupta and B. Park, "A fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 216–229, 2004.
7. K. Y. Gorbunov and V. a. Lyubetsky, "The tree nearest on average to a given set of trees," *Problems of Information Transmission*, vol. 47, pp. 274–288, Oct. 2011.
8. V. a. Lyubetsky and K. Y. Gorbunov, "Fast algorithm to reconstruct a species supertree from a set of protein trees," *Molecular Biology*, vol. 46, no. 1, pp. 161–167, 2012.
9. F. J. Provost and D. N. Hennessy, "Distributed machine learning: scaling up with coarse-grained parallelism," in *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 340–7, Jan. 1994.
10. F. Provost and D. Hennessy, "Scaling up: Distributed machine learning with cooperation," in *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 74–79, 1996.
11. G. J. Williams, *Inducing and Combining Multiple Decision Trees*. PhD thesis, Australian National University, 1990.
12. L. Hall, N. Chawla, and K. Bowyer, "Combining decision trees learned in parallel," *Working Notes of the KDD-97 Workshop on Distributed Data Mining*, pp. 10–15, 1998.
13. L. Hall, N. Chawla, and K. Bowyer, "Decision tree learning on very large data sets," *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, pp. 2579 – 2584, 1998.
14. C. Blake and C. Merz, "UCI Machine Learning Repository," 1998.
15. J. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1992.
16. M. Kuhn, S. Weston, N. Coulter, and R. Quinlan, "C50: C5.0 Decision Trees and Rule-Based Models. R package version 0.1.0-16," 2014.
17. N. Chinchor, "MUC-4 Evaluation Metrics," in *Proceedings of the 4th Message Understanding Conference (MUC4 '92)*, pp. 22–29, Association for Computational Linguistics, 1992.
18. Breiman, Friedman, Olshen, and Stone, *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

A review of recent progress in multi document summarization

Shazia Tabassum[1], Eugenio Oliveira[2]

[1]Ph.D. Student, Faculdade de Engenharia, Universidade does Porto
Porto, Portugal
[2]Professor, Faculdade de Engenharia, Universidade do porto
Porto, Portugal

Abstract. The increase of information available in the form of text, led to the need of extensive research in the area of text summarization. Early the researches in this area started with single document summarization and drove towards multi document summarization**.** We present here a comparative review of the recent progress in the field of multi document summarization. The strengths and weaknesses of the techniques used in the recent researches are highlighted. The state of the art including methods and algorithms on multi document summarization is outlined and discussed. Finally some open research issues are identified.

Keywords: Multi Document summarization, Extraction, Abstraction, Approaches**.**

1 Introduction

One of the major problems being addressed in computer science and informatics from past few years is Big data. A considerable part of the Big data is text oriented. For example Social media, blogs, emails, comments, reviews, news wires etc. The major concerns associated with this information overload are the unlimited text for humans to read or analyze and the limited storage capacity for machines. The present era with huge amounts of unstructured data raises the need for extensive research in the area of text summarization. With the advent of Web 2.0, the Big data would get bigger which implies a strong need for text summarization.

Text Summarization can be seen as an automatic system that makes a précis of text from a single or multiple documents while maintaining the information, meaning, significance and the order of events in the original text. [39] States that, "Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task."

Multi document summarization (MDS) is the task of producing a concise and fluent summary to deliver the major information for a given document set. Multi-document summaries can be used for users to quickly browse document collections, and it has been shown that multi-document summaries can be helpful in information retrieval systems [1].

The summarization task is mainly divided into two categories, extractive summarization and abstractive summarization. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary assuming that these sentences convey the meaning of the whole text. Extraction based summaries produce much less accuracy compared to human made summaries. These methods are easier to apply compared to abstraction based summaries. Abstraction based methods create a compressed version of text conveying the summarized meaning of the original text. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original text [2].

The main goal of this paper is then to overview text summarization different approaches and extract some useful conclusions about them. The rest of the paper is organized as follows: Section 2 outlines the state of the art in text summarization. Section 3 classifies text summaries into different types. Section 4 identifies different approaches followed in previous researches. Section 5 compares the most recent researches in the field. Section 6 presents concluding remarks and scope for future work.

2 State of the Art

Early approaches to text summarization began with the work of [3] in 1950`s. Many approaches have been addressed and many methods have been evaluated since then. Recent approaches used statistical methods such as word frequency, TF-IDF weighting like Sum-Basic [4] [5], sentence position, title relation, and cue-phrases etc. Other approaches take account of semantic associations between words and combine them with those shallow features in the process of sentence similarity. Examples of such approaches are, among others, latent semantic analysis [6], topic signatures [7] and sentence clustering [8].

In recent years, multi-document summarization research has shown increased interest in graph-based approaches [9] [10] [11] [12] [13] [14] and Bayesian topic model based approaches [15] using two-tiered topic model (TTM) in [16] with topic segmentation [17] and topic sum [18]. Others identify the relevance of a sentence by using bigram pseudo sentences for implementing hybrid statistical sentence-extraction [19], rhetoric-based MDS [20] and semantic document concept technique [21] for analyzing grammatical structures in discourses. Recently developed Bayesian collection models incorporated the concept of latent topics into n-gram language models, such as the LDA-HMM model integrating topics and syntax [22], structured topic

model [23], and topical n-grams [24]. [25] Uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster.

On the contrary only few works concentrated on abstractive summarization. Like [37] the author deals with identifying and synthesizing similar elements across related text from a set of multiple documents using natural language text to text generation techniques like content selection, paraphrasing rules, temporal ordering. A fully Abstractive Approach to Guided Summarization is presented by [38] using Information extraction, content selection and generation. Sentence compression [39]. Sentence fusion [40] or sentence revision [41].

3 Types of Summaries

Based on the research work in the field of text summarization, we discuss here the following types of summaries that have been generated.

Generic summaries
Generic summarization tries to extract the most general idea from the original document set without any specified preference in terms of content. For generic summarization, a saliency score is usually assigned to each sentence, the sentences are ranked according to the saliency score, and then the top ranked sentences are selected as the summary based on the ranking result. Recently, both unsupervised and supervised methods have been proposed to analyze the information contained in a document set, and extract highly salient sentences into the summary based on syntactic or statistical features.

Query-Focused Summaries
Query-focused summarization aims at generating a short summary based on a given document set and a given query. The generated summary reflects the condensed information related to the given query within the specified summary length. In query-focused summarization, the information related to a given topic or query should be incorporated into summaries, and the sentences suiting the user's declared information need should be extracted. Many methods for generic summarization can be extended to incorporate the query information.

Update Summaries or incremental summaries
Update summarization is automatically updating summaries as new documents are added to the existing batch of documents. Generating updated summaries as the new documents arrive. Most of existing summarization methods work on a batch of documents and do not consider that documents may arrive in a sequence and the corresponding summaries need to be updated in real time.

Topic Focused Summaries
Topic focused summaries are generated using topic or event based models. A topic model is a type of top-down approach. It considers the same problem based on semantic associations behind the content. In the Bayesian topic model based approaches,

similarity is analyzed using advanced methods with respect to probabilistic distributions of topics [26].

4 Approaches

In order to generate the above types of summaries the approaches followed in the recent works are listed below.

Feature based approach
One of the most common methods used in text summarization field is the feature based method. In the process of identifying important sentences, features influencing the relevance of sentences are determined. Some features that are often considered for sentence selection are word frequency, title words, cue words, sentence location and sentence length [27].

Domain-Specific/Ontology based approach
Generally speaking, ontology is often provided by domain experts [28]. Such ontology provides answers for the questions concerning what entities exist in the domain and how such entities can be related within a hierarchy and subdivided according to similarities and differences among them.

Cognitive based approach
Cognitive psychology is the study of mental processes such as "attention, language use, memory, perception, problem solving, creativity and thinking." [31] Cognitive based approach uses human cognitive factors in reading process. The previous researches in this area have mainly used three cognitive processes, i.e. forget process, recall process and association process for generation of summaries.

Event based approach
Event-based MDS was first proposed by [30], the authors selected sentences based on relevance for one or more sub-events of the topic at hand. Human judges manually determined the sub-events of a topic and assigned to each sentence a relevance score for each sub-event. They show that the algorithm that selects sentences with the highest sum of scores over all sub-events produces the most informative summaries.

Discourse based approach
Discourse is an organic structure. Different parts of discourse bear different functions, and have complex relationships among them. Automatic summarization based on discourse attempts to analyze the structural features of discourse to identify the main content of the article. Currently, automatic summarization based on discourse has five main research topics: rhetorical structure analysis, pragmatic analysis, lexical chain, relationship map and latent semantic analysis [31].

Table. I The comparison of most recent researches on MDS.

NAME	*1.A MDS system based on statistics and linguistic treatment (2014) [11]*	*2.A novel contextual topic model for multi-document summarization (2015) [26]*
CATEGORY	*Extractive*	*Extractive*
TYPE	*Generic*	*Topic focused*
APPROACH	*Statistics, Linguistics and machine learning*	*Contextual topic model based approach*
METHOD	*Graph based clustering algorithm*	*Bayesian topic model*
HIGHLIGHTS	*To deal with multi document issues such as redundancy and problem diversity*	*A model that can capture both the hierarchies and the word dependencies over latent topics*
LIMITATIONS/IMPOVEMENTS	*The limitation of the approach is the problem of sentence ordering, as the system tries to find relevant sentences in groups of different topics.*	*The model has to take longer time to be trained under a larger data set in order to keep certain level of accuracy in prediction. It is poor to cover co-occurrence among multiple words. Other limitations include the problem of sentence coherence and a lack of online settings for stream text because the model has been trained using a large data corpus before summarizing documents..*
EVALUATION ROUGE 2	*200 word sumary: Recall- 62% .400 word summary: Recall 53%*	*DUC(2006) Recall 0.0986*

5. *Event graphs for information retrieval and multi-document summarization (2014) [12]*	*4.FoDoSu:Multi-document summarization exploiting semantic analysis based on social Folksonomy (2015) [32]*	*3.Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization(2014) [9]*
Extractive	*Extractive*	*Extractive*
Event focused	*Generic*	*Query based*
Graphs, machine learning and rule based extraction methods	*Using statistical and linguistic methods*	*Graph model using matrix factorization*
Event graphs, Logistic regression classifier, argument extraction, Temporal relation extraction	*HITS algorithm and semantic analysis.*	*Weighted archetypal analysis*
The contribution of this article is a novel event-centered document representation that accounts for the semantics of events. The other contribution of this article is a novel event-centered mds.	*To analyze the relationship between the semantics of the words in Web documents.*	*1.To incorporate query information in its own nature of an archetypal analysis.* *2. To increase variability and diversity of the produced query based summary.*
To improve the extraction of temporal relations between events which would yield further performance improvements in event-centered retrieval and summarization. The second improvement would be to enrich event graphs with other relations that can hold between events,such as causality, entailment,and atiotemporal contai nment..	*To improve methods for analyzing the semantics of words that is difficult to analyze, such as proper nouns and newly-coined words.*	*In future work WordNet could be used to calculate the semantic similarity between sentences by using the synonyms sets of their component terms;Another possible enhancement can be reached by introducing the multi-layered graph model that emphasizes not only the sentence to sentence and sentence to terms relations but also the influence of the under sentence and above term level relations, such as n-grams, phrases and semantic role arguments levels.*
DUC 2002,2004 Recall 0.116 ,0.107	*Tac 2008, tac 2009 Recall 0.06853*	*DUC(2006) Recall - 0.0917*

8.Cognitive based MDS (2014) [34]	*7.An empirical study on ontology based multi document summarization in disaster management (2014) [28]*	*6. Multi document summarization based on news components using fuzzy cross – document relations (2014) [33]*
Extractive	*Extractive*	*Extractive*
Topic focused	*Generic and Query focused*	*Generic*
Cognitive based	*Ontology based using statistics and machine learning*	*Feature based approach using evolutionary algorithm, Fuzzy logic, machine learning*
IRatio, GWI(global word impression)LWI(local word impression)	*TF, TF-IDF, TF-ICF, Concept hierarchy and clustering*	*Genetic algorithm, Case based reasoning, classification and fuzzy scoring*
Proposed inter-document recall process and forget process in the scanning mechanism.	*Introduction of domain specific ontology in disaster management for generic and query focused summaries.*	*Introduction of a multi document summarization model by taking into account the generic components of news story. The study further investigates the utility of cross-document relations (CST relations) to identify highly relevant sentences to be included in the summary.*
To employ sophisticated methods on understanding the semantics and redundancy of sentences to improve the quality of summary.	*To utilize the hierarchical correlations in the ontology to further improve the quality of the summary. To employ information extraction techniques to further improve summarization results.*	*To explore how natural language processing techniques can be employed to connect semantic concepts with news components. To study the utility of cross-document relations identified from un-annotated text documents to generate better summaries by treating issues related to multi document such as contradictions and historical information.*
UCI 2007 Recall 0.12288	*Hurricane dataset Recall 0.30160*	*DUC 2002 Recall - 0.1280*

11.Exploring actor–object relationships for query-focused MDS(2014) [36]	*10. SRRank: Leveraging Semantic Roles for Extractive MDS(2014) [1]*	*9.Incremental MDS: An Incremental Clustering Based Approach (2014) [35]*
Extractive	*Extractive*	-
User Focused	*Topic Focused*	*Incremental/ Update*
User based, Feature based , Machine learning, Actor Object relationship (AOR)	*Graph based ranking algorithm*	*Machine learning, Clustering*
Back propagation on neural network,	*Semantic parsing SRRank*	*Clustering*
Combines ensemble summarizing system and AOR to generate summaries	*Proposes a novel graph-based sentence ranking algorithm SRRank to incorporate the semantic role information into the graph-based ranking algorithm.*	*Study of order dependency on clustered documents*
Plan to examine further the effectiveness of exploiting other types of patterns resulting from using dependency parsers in a user-based summarization system and investigate their effectiveness.	*making use of deep semantic information instead of shallow semantic information for further improving multi-document summarization.*	*For maintaining the summaries stable, there must be some mechanism that keeps the extracted important sentences in the order irrespective of the order in which the input documents are handled by the program..*
DUC 2006 & 2007 Recall 0.933, 0.1219	*UCI 2006 & 2007 Recall 0.9044 & 0.956*	-

5 Comparison of most recent Researches

In this section we present a comparative study of the researches on multi document text summarization techniques from previous year. Recent researches have concentrated on different approaches discussed in the previous section. Table I highlights the comparative points between those techniques. We have pointed out different types of summaries generated by using different approaches and methods. We have also brought out the limitations or improvements suggested for those approaches. Last column reports on the evaluation results using ROUGE set of metrics.

6 Conclusion and future work

In this paper we have clearly classified text summarization into different types, following different approaches and using different methods. We have also compared the most recent researches in the area of MDS. Referring to the work above, we would say that there are very few works using abstraction while most of the summaries use extractive approaches. The research work in text summarization is expanding with the implementation of methods and methodologies from various fields like cognitive psychology, evolutionary algorithms, discourses etc. By analyzing the results of the recent research, we found that the works using these fields have outperformed the previous methods, but are still far from human generated summaries.

Some limitations in previous research works help us to identify few research issues in MDS like, problem of sentence ordering, redundancy of sentences, enriching graphs with semantic relationship between sentences and documents, improving feature extraction methods, understanding the human way of summarizing, improving coherence in summaries.

These challenges can be seen as relevant motivation and possible guidelines for future research topics in the area of MDS.

References

1. Yan S, Wan X (2014) SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization. IEEE/ACM Transactions on audio, speech, and language processing, Vol.22, No.12.
2. Rafeeq Al-Hashemi (June 2010) Text Summarization Extraction System (TSES) Using Extracted Keywords. International Arab Journal of e-Technology, Vol. 1, No. 4
3. Luhn HP (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.
4. Nenkova A & McKeown K (2012) A survey of text summarization techniques. In Mining text data , US: Springer, pp 43–76.

5. Vanderwende L, Suzuki H, Brockett, C & Nenkova A (2007) Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing & Management, 43(6), 1606–1618.
6. Gong Y & Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 19–25.
7. Lin CY & Hovy E (2000) The automated acquisition of topic signatures for text summarization. In Proceedings of the international conference on computational linguistic. pp 495–501.
8. He R, Qin B & Liu T (2012) A novel approach to update summarization using evolutionary manifold-ranking and spectral clustering. Expert Systems with Applications, 39(3), 2375–2384.
9. Canhasi E & Kononenko I (2014) Weighted archetypal analysis of the multi element graph for query-focused multi-document summarization. Expert Systems with Applications, 41(2), 535–543.
10. Ferreira R, de Souza Cabral L, Lins RD, Pereira e Silva G, Freitas F, Cavalcanti GD et al (2013) Assessing sentence scoring techniques for extractive text summarization. Expert Systems with Applications, 40(14), 5755–57
11. Ferreira R, de Souza Cabral L, Freitas F, Lins R D, de França Silva G et al (2014) A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications, 41(13), 5780–5787.
12. Glavaš G & Šnajder J (2014) Event graphs for information retrieval and multi-document summarization. Expert Systems with Applications, 41(15), 6904–6916.
13. Mendoza M, Bonilla S, Noguera C, Cobos C & León E (2014) Extractive single document summarization based on genetic operators and guided local search. Expert Systems with Applications, 41(9), 4158–4169.
14. Zhao L, Wu L & Huang X (2009) Using query expansion in graph-based approach for query-focused multi-document summarization. Information Processing & Management, 45(1), 35–41.
15. Daumé III H & Marcu, D (2006) Bayesian query-focused summarization. In Proceedings of the conference of the association for computational linguistics (ACL) and 44th annual meeting of the ACL, Sydney, pp 305–312.
16. Celikyilmaz A & Hakkani-Tur D (2011) Discovery of topically coherent sentences for extractive summarization. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Vol. 1, pp. 491–499.
17. Eisenstein J & Barzilay R (2008) Bayesian unsupervised topic segmentation. In Proceedings of the conference on empirical methods in natural language processing, Oct 25–27, Honolulu, Hawaii.
18. Haghighi A & Vanderwende L (2009) Exploring content models for multi document summarization. In Proceedings of human language technologies: The annual conference of the North American chapter of the association for computational linguistics, Boulder, Colorado, pp. 362–370.
19. Ko Y & Seo J (2008) An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Pattern Recognition Letters, 29(9), 1366–1371.
20. Atkinson J & Munoz R (2013) Rhetorics-based multi-document summarization. Expert Systems with Applications, 40(11), 4346–4352.
21. Ye S, Chua T S, Kan M Y & Qiu L (2007) Document concept lattice for text understanding and summarization. Information Processing & Management, 43(6), 1643–1662.

22. Griffiths T, Steyvers M, Blei D & Tenenbaum J (2005) Integrating topics and syntax. Advances in Neural Information Processing Systems, 17.
23. Wallach H M (2006) Topic modeling: Beyond bag-of-words. In Proceedings of the 23rd international conference on machine learning , ACM, pp. 977–984.
24. Wang X, McCallum & Wei X (2007) Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th IEEE international conference on data mining , pp 697–70.
25. Radev DR, Jing H, Stys M and Tam D (2004) Centroid-based summarization of multiple documents. Information Processing and Management, 40, 16-17.
26. Yang G, Wenb D, Kinshuk, Chen N, Sutinen E (2014) A novel contextual topic model for multi-document summarization. Expert Systems with Applications, 42, 1340–1352.
27. Gupta V, Lehal GS (2010) A survey of text summarization extractive techniques, J. Emerg. Technol. Web Intell. 2 258–268.
28. Wu K, Li L, Li J, Li T (2013) Ontology-enriched multi-document summarization in disaster management using sub modular function. Information Sciences 224, 118–129.
29. American Psychological Association (2013). Glossary of psychological terms. Apa.org. Retrieved 2014-08-13.
30. Daniel N, Radev D & Allison T (2003) Sub-event based multi-document summarization. In Proceedings of the HLT-NAACL 03 Workshop on Text Summarization, Association for Computational Linguistics, Vol. 5, pp 9–16.
31. Wang S, Li W, Wang F, Deng H (2010) A Survey on Automatic Summarization. International Forum on Information Technology and Applications.
32. Heu J, Qasim I, Lee D (2015) FoDoSu: Multi-document summarization exploiting semantic analysis based on social Folksonomy. Information Processing and Management 51, 212–225.
33. JayaKumar Y, Salim N, Abuobied A, Albaham A T (2014) Multi document summarization based on news components using fuzzy cross-document relations. Applied Soft Computing 21, 265–279.
34. Chen J, Li W (2013) Cognitive-based Multi-Document Summarization Approach. Ninth International Conference on Semantics, Knowledge and Grids.
35. Johney John, Asharaf S (2014) Incremental Multi-Document Summarization: An Incremental Clustering Based Approach. International Conference on Data Science & Engineering (ICDSE).
36. Valizadeh M & Brazdil P (2014) Exploring actor–object relationships for query-focused multi-document summarization. Soft Computing. doi:10.1007/s00500-014-1471-x.
37. Regina Barzilay and Kathleen R. McKeown (2005) Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3):297-328.
38. Genest PE and Lapalme G (2012) Fully abstractive approach to guided summarization. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. Jeju Island, Korea, Association for Computational Linguistics: 354-358.
39. Cohn T and Lapata M (2009) Sentence compression as tree transduction. J. Artif. Int. Res. 34(1): 637-674.
40. Barzilay R et al (1999) Information fusion in the context of multi-document summarization. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland, Association for Computational Linguistics, pp 550-557.

41. Tanaka H et al (2009) Syntax-driven sentence revision for broadcast news summarization. Proceedings of the 2009 Workshop on Language Generation and Summarization. Suntec, Singapore, Association for Computational Linguistics**,** pp 39-47.

Text Mining Scientific Articles using the R Language

Carlos A.S.J. Gulo[1,2] and Thiago R.P.M. Rúbio[1,3]

[1] Faculdade de Engenharia da Universidade do Porto (FEUP)
Departamento de Engenharia Informática (DEI)
[2] LAETA-Laboratório Associado de Energia, Transporte e Aeronáutica (FEUP)
PIXEL Research Group - UNEMAT/Brazil (http://goo.gl/tcg6S7)
Web: http://lattes.cnpq.br/0062065110639984 - sander@unemat.br
[3] LIACC Research Group (FEUP)
reis.thiago@fe.up.pt

Abstract. The aim of this study is to develop a solution for text mining scientific articles using the R language in the "Knowledge Extraction and Machine Learning" course. Automatic text summary of papers is a challenging problem whose approach would allow researchers to browse large article collections and quickly view highlights and drill down for details. The proposed solution is based in social network analysis, topic models and bipartite graph approaches. Our method defines a bipartite graph between documents and topics built using the Latent Dirichlet Allocation topic model. The topics are connected to generate a network of topics, which is converted to bipartite graph, using topics collected in the same document. Hence, it is revealed to be a very promising technique for providing insights about summarizing scientific article collections.

Keywords: Text Mining, Topic Model, Topic Network, Systematic Literature Review

1 Introduction

With the overwhelming amount of textual information presented in scientific literature, there is a need for effective automated processing that can help scientists to locate, gather and make use of knowledge encoded in literature that is available electronically. Although a great deal of crucial scientific information is stored in databases, the most relevant and useful information is still represented in domain literature.

The literature review process consists of: to locate, appraise, and synthesize the best-available empirical evidence to answer specific research questions. An ideal literature search would retrieve all relevant papers for inclusion and exclude all irrelevant papers. However, previous research have demonstrated a number of studies that are not fully indexed, as well as a number that are indexed incorrectly [7,15,8].

The purpose of this paper is to highlight text mining techniques as a support to identify the relevant literature from a CSV (Comma Separated Value)

collection searched in different journal repositories. Data set will be analyzed quantitatively in order to obtain a systematic review literature process involving the research domain: *high performance computing as support to computer aid diagnostic systems*. This research domain is the first author's scientific field of interest.

Text Mining is a common process of extracting relevant information using a set of documents. Text Mining provides basic preprocessing methods, such as identification, extraction of representative characteristics, and advanced operations as identifying complex patterns [11,1,5]. Document classification is a task that consists of assigning a text to one or more categories: the name of its class of subject, and main topics. This paper only addresses the summarization of *Abstracts*. Researchers are interested in the number of times certain keywords associated with specific content appear in each document.

This paper is organized as follows: in the next section is described a summary about text classification. The experiments performed using the R code and the results obtained with the sets of scientific articles considered in the automatic text summary and text classification, are discussed in Section 3, which is followed by the concluding remarks in Section 4.

2 Related Work

This section summarizes some achievements on text classification from various pieces of the literature. In general, text classification is a problem divided into nine steps. Those steps include data collection, text processing, data division, feature extraction, feature selection, data representation, classifier training, applying a classification model, and performance evaluation [12,18].

- Data Collection: In text classification, the first step is collecting data. The sample data are texts that belong to a limited scientific domain, i.e., "high performance computing as support to medical image processing" [17]. Each sample text must be labeled with one or more tags indicating a *label* to a certain class.
- Text preprocessing: Actually is preprocessing a trial to improve text classification by removing worthless information. It may include removal of punctuation, stop words (any prepositions and pronouns), and numbers [18,9]. In the context of this paper, we consider root extraction and word stemming as part of the feature extraction step [12], which will be discussed in the Feature extraction item.
- Data division: Next step divides the data into two parts, training data and testing data. Based on training data, the classification algorithm will be trained to produce a classification model. The testing data will be used to validate the performance of the resulting classification model. There is no ideal ratio of training data to testing data. The text classification experiments presented have been used 25% for training and 75% for testing. The classification performance is the average performance of implemented classification models[19,12].

– Feature extraction: Texts are characterized by features that: *a)* are not related to the content of the text, such as author gender, author name, and others; and *b)* reflect the text content, such as lexical items and grammatical categories. Considering single words, the simplest of lexical features, as a representation feature in text classification has proven effective for a number of applications. The result of this step is a list of features and their corresponding frequency in the training data set [18].

– Feature selection: The result of the feature extraction step is a long collection of features, however, not all of these features are good for classification for many reasons: first, some classification algorithms are negatively affected when using many features due to what is called "curse of dimensionality" next, the over-fitting problem may occur when the classification algorithm is trained in all features and finally some other features are common in most of the classes. To solve these problems, many methods were proposed to select the most representative features for each class in the training data set. In this paper, the most frequently used methods have been Chi Squared (CHI), term frequency (TF), document frequency (DF) and their variations. Other than statistical ranking, features with higher frequency were used. Word stems are also used as feature selections where words with the same stem are considered as one feature [19,18,17].

– Data representation: The results obtained from the previous step are represented in matrix format, and will be used by the classification algorithm. Usually, the data are in matrix format with n rows and m columns wherein the columns correspond to the selected feature, and the rows correspond to the texts in the training data. Weighting methods, such as term frequency inverse document frequency (TFIDF) and term frequency (TF) are used to compute the value of each cell in this matrix, which represents the weight of the feature in the text [9].

– Classifier training: The classification algorithm is trained using the training matrix that contains the selected features and their corresponding weights in each text of the training data. Support Vector Machine (SVM) and Naïve Bayes (NB) are the classical machine learning algorithms that have been the most used in text classification [10,1]. The result is a classification model to be tested by means of the testing data. The same weighting methods and the same features extracted from the training data will be used to test the classification model [16,19].

– Classification model evaluation: Evaluation techniques are assessed to estimate future performance by measures such as accuracy, recall, precision, and f-measure, and to maximize empirical results [17].

Table 1. Total of articles searched in journal repositories.

Repositories	Publication	
	Searched Queries	Papers
ACM Portal	("medical image") and ("high performance computing" or "parallel computing" or "parallel programming")	1
Engineering Village	(((((("medical imag*") WN KY) AND (("high performance comput*") WN KY)) OR (("parallel comput*") WN KY)) OR (("parallel programm*") WN KY)), Journal article only, English only	19
IEEE Xplore	((medical imag*) AND (("high performance comput*" OR "parallel programm*") OR "parallel comput*"))	69
ScienceDirect	"medical image" AND ("high performance computing" OR "parallel computing" OR "parallel programming")[Journals(Computer Science,Engineering)]	390
Web of Science	TOPIC: ("medical imag*") AND TOPIC: ("high performance comput*") AND TOPIC: ("parallel")	27
Total		506

3 Experiments and Discussion

This section describes the infrastructure used to perform the experiments and also illustrates and discusses the results obtained. Data set[4] used in experiments were collected from repositories showed in Table 1, and composed by 7 variables (*id*, *Title*, *Journal*, *Year*, *Abstract*, *Keywords* and *Recommend*) and 494 observations (after removing duplicated records).

We are interested in what the characteristics are *Abstract* that tend to group the article in a specific topic, and in future work recommend the prioritized observations based on high scores of topics. The analyzed variable is text data, the *Abstract*, and its unstructured data. Unstructured data has variable length, one observation contains an academic text, it has variable spelling using singular and plural forms of words, punctuation and other non alphanumeric characters, and the contents are not predefined to adhere to a set of values - it can be on a variety of topics [6,3].

To create useful data, unstructured text data should be converted into structured data for further processing. The preprocessing step, described in Section 3.1, involves extraction of words from the data and removal of punctuation and spaces, eliminates articles and other words that we are not interested in, replaces synonyms, plural and other variants of words with a single term and finally, makes the structured data, which is a table where each word becomes a variable with a numeric value for each record.

[4] Project code and data set is available in https://github.com/carlosalexsander/ECAC_Project

The test infrastructure used was composed of the RStudio development suite, available to download through the RStudio website [5]. A graphic card GeForce GT 540 (NVIDIA) with 192 CUDA cores and 2GB of GDRAM was used with a portable computer equipped with an Intel(R) Core(TM) i7-2630QM 2.0 GHz, 8GB of RAM (DDR3 1333 MHz), Linux Debian Wheezy (64 bits) operating system.

3.1 Results and Discussion

In the first step, it was necessary to install and load the R package Text Mining *tm* to process text documents. Once we have a corpus, the next step was to modify the documents in it, e.g., making everything lowercase, reducing words to their stem, removing numbers, removing punctuation, and removing common English stop-words. All this functionality is named a transformation concept, illustrated in Fig. 1. In general, all transformation work is done in all elements of the corpus applying the *tm_map* function.

```
toSpace <- content_transformer(function(x,pattern) gsub(pattern, " ", x))
corpus.m <- tm_map(paper.corpus, toSpace, "/|@|\\|")
corpus.m <- tm_map(corpus.m, content_transformer(tolower))
corpus.m <- tm_map(corpus.m, removePunctuation)
removeUnicode <- function(x) stri_replace_all_regex(x,"[^\x20-\x7E]","")
corpus.m <- tm_map(corpus.m, content_transformer(removeUnicode))
corpus.m <- tm_map(corpus.m, removeNumbers)
corpus.m <- tm_map(corpus.m, removeWords, stopwords("english"))
corpus.m <- tm_map(corpus.m, removeWords, c("using", "used", "propose", "can
    ", "also"))
library(SnowballC)
corpus.m <- tm_map(corpus.m, stemDocument, language = "english")
corpus.m <- tm_map(corpus.m, stripWhitespace)
corpus.dtm <- DocumentTermMatrix(corpus.m, control = list(minWordLength = 3,
      weighting = function(x) weightTFIDF(x, normalize = FALSE)))
```

Fig. 1. Preprocessing text documents applying function *tm_map*().

For many methods of text analysis, specifically the so called "bag-of-word" approaches, we created a common data structure for the corpus (Document Term Matrix - DTM) [13,17]. This is a matrix in which the rows represent documents and columns represent terms. The values represent how often each word occurred in each document. Not all terms are equally informative of the underlying semantic structures of texts, and some terms are rather useless for this purpose. We used the *term.statistics* function, created in order to produce text statistics, for instance, the most common words in the text, illustrated in Fig. 2.

For interpretation and computational purposes it is worthwhile to delete some less useful words from the DTM before fitting a model [14]. We can now filter out

[5] http://www.rstudio.com/products/rstudio/download/

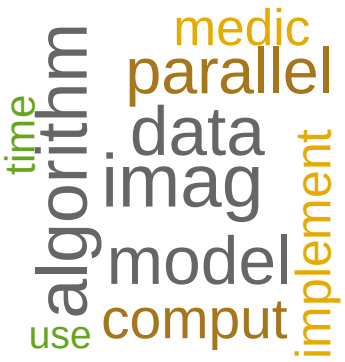


Fig. 2. Most frequented words in corpus represented by wordcloud.

words based on this information, select only the terms with the highest TFIDF score, and after apply the function *removeSparseTerms(DTM, S)* to retain the fewer (but more common) terms. The *sparse* argument value used was 0.75 to retain more words for classification than with a smaller *sparse* value, and those words were used in dictionary-based approach for term identification. When the removal of sparse terms function are applied, the number of terms is reduced from 4851 to 22.

TFIDF, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in text mining. The TFIDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps control the fact that some words are generally more common than others. Variations of the TFIDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query[14]. TFIDF can be successfully used for stop-words filtering in various subject fields including text summary and classification. One of the simplest ranking functions is computed by summing the TFIDF for each query term; many more sophisticated ranking functions are variants of this simple model[10]. The TFIDF formula is:

$$TFIDF(i) = \frac{Frequency(i) * N}{df(i)}, \tag{1}$$

where df is the frequency of word (i) in all documents, and N is the number of words in the record/document. An interesting technique to use on a document-term matrix is the Latent Dirichlet Allocation (LDA) topic modeling. Topic modeling are statistical methods, essentially used to analyze the words of the original documents to discover the topics that run through them and how those

topics are connected to each other [6,2]. Before fitting the topic model, coded in Fig. 3, it is best to reduce the vocabulary by selecting only informative words.

```
best.model <- lapply(seq(2, 10, by = 1), function(d){LDA(term.tfidf.df, d)})
best.model.logLik <- as.data.frame(as.matrix(lapply(best.model, logLik)))
best.model.logLik.df <- data.frame(topics=c(2:10), LL = as.numeric(as.matrix
    (best.model.logLik)))
best.model.logLik.df.sort <- best.model.logLik.df[order(-best.model.logLik.
    df$LL), ]
best.model.logLik.df.sort
ntop <- best.model.logLik.df.sort[1,]$topics
lda <- LDA(term.tfidf.df, ntop)
```

Fig. 3. Code to produce a list of likelihood for each model and to optimize the LDA model.

Weights were used to determine the most efficient way. The treatment is mathematically straightforward. [16,4,3], The best model fitted is depicted by the distribution of this likelihood by topic in Fig. 4. The number of topics with the highest log likelihood is −308886.3 and 10. These topics gave the best fit for the present data.

A different plot to see the final result is the topic network. It's created by multiplying the document-topic matrix transpose with itself by an adjacency matrix. Fig. 5 shows the topic network, with topic nodes labeled by the three most probable terms in the corresponding topic distribution. In the context of this study, we define *topic* to be a distribution over a fixed vocabulary, i.e. *image* topic has words about image with high probability and the *image processing* topic has words about image processing with high probability.

4 Conclusions

In this work we studied how to use R language to text mining, and the great importance of a good preprocessing of the data in order to obtain good quality results in the text classification step. In general, building a large amount of labeled training data for text classification is a labor-intensive and time-consuming task. That was the main problem during the classifier development, and it is defined as future work. We found that the simplest document representation was at least as good as representations involving more complicated syntactic and morphological analysis. Topic networks may be useful in analyzing documents collections.

5 Acknowledgments

The first author thanks Universidade do Estado de Mato Grosso (UNEMAT - Brazil), for the support given. The work of Thiago, has been funded through a

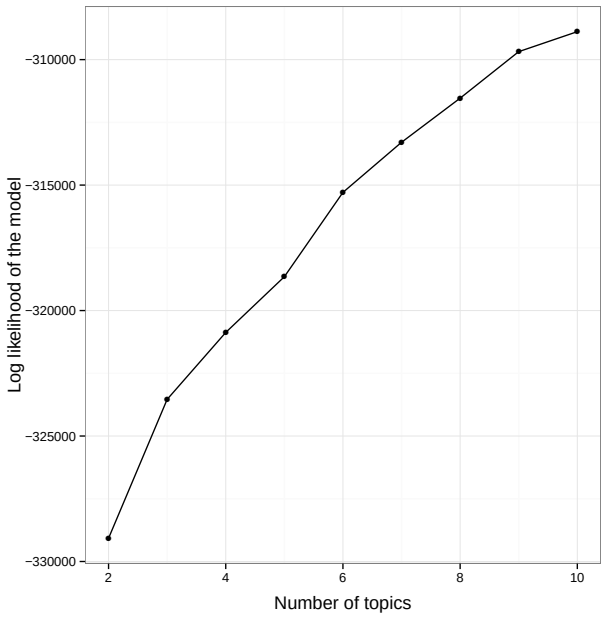


Fig. 4. The distribution of likelihood by topic.

IBRASIL Grant. IBRASIL is a Full Doctorate programme selected under Erasmus Mundus, Action 2 STRAND 1, Lot 16 and coordinated by University of Lille.

References

1. Aphinyanaphongs Y FAU Aphinyanaphongs, Y., Aliferis C FAU Aliferis, C.: Text categorization models for retrieval of high quality articles in internal medicine. In: AMIA Annual Symposium Proceedings. pp. 31–35. No. 1942-597X (Electronic) (2003), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480096/
2. Blei, D.M.: Surveying a suite of algorithms that offer a solution to managing large document archives. Communication of the ACM 55(4), 77–84 (April 2012)
3. Blei, D.M., Lafferty, J.D.: Topic models (2009)
4. Blei, D., Ng., A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
5. Casualty Actuarial Society E-Forum: Text Mining Handbook (March 2010), http://www.casact.org/pubs/forum/10spforum/Francis_Flynn.pdf
6. Cohen AM FAU Cohen, A.M., Ambert K FAU Ambert, K., McDonagh M FAU McDonagh, M.: Cross-topic learning for work prioritization in systematic review creation and update. In: Journal of the American Medical Informatics Association?:

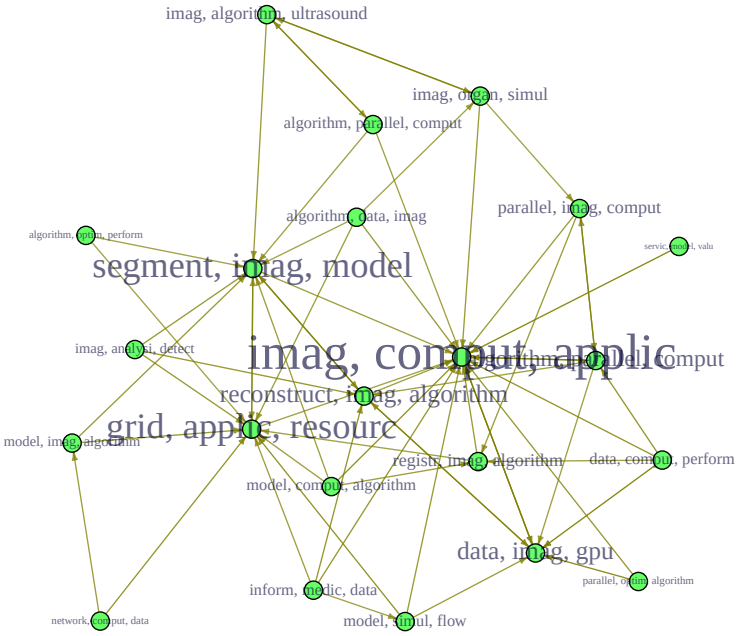


Fig. 5. Final result plotted in a Topic Network.

JAMIA. pp. 690–704. No. 1527-974X (Electronic) (2009), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2744720/

7. Cooper, H.M.: The structure of knowledge synthesis, Knowledge in Society, vol. 1 (1988)
8. Edinger T FAU Edinger, T., Cohen AM FAU Cohen, A.M.: A large-scale analysis of the reasons given for excluding articles that are retrieved by literature search during systematic review. In: AMIA Annual Symposium Proceedings. pp. 379–387. No. 1942-597X (Electronic) (Nov 2013), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900186/
9. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. Journal of Statistical Software 25(5), 1–54 (3 2008), http://www.jstatsoft.org/v25/i05
10. Hotho, A., Nürnberger, A., Paab, G.: A brief survey of text mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology (2005)

11. Kao, A., Poteet, S.R. (eds.): Natural Language Processing and Text Mining. No. ISBN 1-84628-175-X, Springer (2007)
12. Khorsheed, M., Al-Thubaity, A.: Comparative evaluation of text classification techniques using a large diverse arabic dataset. Language Resources and Evaluation 47(2), 513–538 (2013), http://dx.doi.org/10.1007/s10579-013-9221-8
13. Lebanon, G., Mao, Y., Dillon, J.: The locally weighted bag of words framework for document representation. J. Mach. Learn. Res. 8, 2405–2441 (Dec 2007), http://dl.acm.org/citation.cfm?id=1314498.1314576
14. Munzert, S., Rubba, C., Meibner, P., Nyhuis, D.: Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. No. ISBN 978-1-118-83481-7, John Wiley & Sons, Ltd, 1st edn. (2015)
15. Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. Sprouts: Working Papers on Information Systems 10(26) (2010), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1954824
16. Reed, C.: Latent Dirichlet Allocation: A Student Companion (2012)
17. Weiss, S.M., Indurkhya, N., Zhang, T.: Fundamentals of Predictive Text Mining. No. e-ISBN 978-1-84996-226-1, Springer (2010)
18. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J.: Text Mining: Predictive Methods for Analyzing Unstructured Information. No. ISBN 0-387-95433-3, Springer (2005)
19. Zhao, Y.: R and Data Mining: Examples and Case Studies. Academic Press (2013), http://www.sciencedirect.com/science/article/pii/B9780123969637000015

Characterizing Developers' Rework on GitHub Open Source Projects

Thiago R.P.M. Rúbio and Carlos A. S. J. Gulo

LIACC / DEI, Faculdade de Engenharia, Universidade do Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal,
{reis.thiago,prodei1300766}@fe.up.pt

Abstract. Identify Rework on Open Source Software (OSS) projects is a challenging, complex and still open issue. We believe that in OSS environment, team members should introduce less rework than volunteers, due to their knowledge about the project, but how to characterize and classify developers' actions as rework sources? In this paper we presented a novel approach for classifying commits by analysing data from open source repositories, named in GitHub. We constructed our classifier following a data mining methodology and comparing the performance of three algorithms: Decision Trees, Naive Bayes and K-NN. For testing our model, we applied data of 3311 commits from the Subversion project. Results show that K-NN outperforms the others in this task and the model can predict whether a commit is a rework source or not with 70% of accuracy. Our training data depends much on the relation commits-issues and and the collecting methodology should be improved. We envisage to create a more general classification model by using a bigger dataset including data from many other projects.

Keywords: Rework, Data Mining, Classification, Open Source, OSS, GitHub

1 Introduction

Software development is facing a big change. Traditionally, companies have teams that work under a certain budget to achieve milestone objectives. On the other side, the Open Source Software (OSS) has its importance growing in the last years and is changing the rules of this game [1]. The Open Source initiative is based on free code which any other developer can contribute with the project by free will.

One big problem in this scenario is the work introduced by issues not solved consistently or bugs created by developers. This work was not planned in project specification and is called Rework. Rework can be defined as the time and effort considered in trying to fix some bug or solving an issue that was already considered solved. It implies big costs because of the incremental nature of software. Earlier a problem is found, less is the cost to be solved [2].

In OSS software, finding rework sources could help to discover developers' behavior and create a profile for each of them, making it easy to select the best

team members or even select those who need help to improve their programming. In fact, it helps to reduce costs, improving the quality of code. Identifying and reducing rework is fundamental for decreasing the cost of development and maintenance of the software [2]. Companies are also interested in this because they could benefit mixing their traditional techniques with ours and reduce even more their expenses with rework. Traditional development has a much more controlled environment and guided by project budgets, they have their own processes for dealing with rework. In OSS software the main difficulty is the lack of knowledge about the quality of code coming from different developers with very different profiles [3]. The impact of rework in OSS software should be studied as a way to know and compare how open source projects are affected by rework.

In this paper we present a data mining approach to create developers' profiles and classify their actions based on the observation of real open source projects development data. We have created a general model, using well-known algorithms and evaluated their performance to obtain a good prediction.

Using a data mining methodology we have compared several algorithms and selected the best model for the classification task, expecting to improve the quality and the reliability of the predictions. We believe that our model can be a good ally on identifying rework sources in projects and helping developers to manage new code. This work had three main goals (i) use data mining tools and methodology to create a working model for classifying real OSS data; (ii) compare the performance of classification using different algorithms and select the best model (iii) analyse the model's predictions and evaluate developers' profiles in OSS projects.

Analysing a real world project's source repository is a big challenge, but also the results are very interesting [4]. Outcomes from our model included the characteristics of the roles members and volunteers and their relation to the creation of rework. Although the type of developer was very important, we wanted to discover more about the relation between rework and other features on developer's actions. Our results proved to be very surprising. We have found that members do also introduce a considerable amount of rework, but volunteers work in simpler tasks.

The rest of this paper is structured as follows: In Section 2 we present the Open Source platforms scenario. Section 3 discusses the problem of based on reviewing the literature introducing rework in software projects and how establishing a developer profile could help reducing efforts in tackling software bugs. We also describe the GitHub environment as an example of OSS environment. In Section 4 we describe the experimental design of our classification model. Section 5 presents the experimental evaluation of the model. We discuss the findings of this work and point lines of future research in Section 6.

2 Related Work

Open Source Software (OSS) is a trend in software development. Since late 1990, an uncountable number of projects have grown in this environment proving

it can be successful and and profitable [5]. Research interest in OSS is very diversified. In some way, OSS platforms represent a social networking model between developers and customers, useful information for analysing software marketing and customers preferences [6]. In other way, people enrolled with OSS highlight the relationship between the customers as volunteers and the satisfaction with the product [4]. Big software companies (e.g., Microsoft, Google GNOME, Linux, etc.) are also focusing OSS projects in the last years motivated by the collaboration between volunteers and developers, which can reduce costs, spread the technology and help developing more user-guided tools.

Given its free nature, OSS create a difficulty in managing resources, planning and delivering projects. As said in [7], resource allocation and budgeting is even a harder challenge. The cost of the development is an important factor for the success of a project [7] and here we find the rework cost. As explained in section 1, rework consists on the effort and consequently monetary cost of trying to fix something that was already considered a closed issue. Rework is, in fact, a big problem in software engineering, consuming big part of the project budget (40% up to 70%) [8]. Introducing rework could be explained by human problems in project management like communication, formation and work conditions [8] and the Industry believes that great part of rework could be early identified and avoided, but until now not much attention has been paid in studying rework.

We point out the relation between rework and code quality, measured regarding the actions performed by developers, named commits. Commits are the registries of the work and have the information of all modifications in previous files and new code related to some issue discussed by the team. Commits provide us the information about how work was done and that could be used for predicting if a commit represent or not a rework source.

3 Problem Statement

In a software project, all commits must be indexed by an issue and usually, software development includes two types of tools for managing this: a SCM (Software Configuration Management) tool and an Issue Tracker. SCM tools are responsible for storing all the code produced and all the information about modifications made in the development process (commits). Issue Trackers are responsible for project management, storing the issues and their relation with commits. In Issue Trackers we can see the history of the project and how commits effectively concludes each activity needed.

Our work was to merge the information in this two sources and analyse the data of a real project. We choose the GitHub [1]platform because of its great importance in OSS environment. Known as the largest open source community, GitHub is one of the most important code hosting sites and has a successful social component useful for analysing developers' behavior [9,10]. Based on the Git version control system, GitHub hosts all kinds of open source projects, with

[1] GitHub.com

different sizes and programming languages. Big companies have a very important place in GitHub as they carry large projects and high numbers in the platform.

We have selected the Apache Subversion (SVN) [2] project by its relevance and importance in GitHub platform. A software versioning and revision control system distributed as free software under the Apache License. Recognized as the sole leader in the Standalone SCM category, Subversion has in GitHub about 50000 commits divided into 768 branchs and 238 defined milestones. Fifteen developers are considered members of the actual development team and the average development productivity in Subversion is about 300 commits and 140 files modified by 8000 lines of code per month. Subversion also uses the Tigris[3] platform as the Issue Tracker system and have five types of issues: DEFECT, CLOSED, NEW, REOPENED, and STARTED, in a total of approximate 2690 activities. This system provides us the information regarding which commits are related to these issues and their committers.

4 Model and Experimental Setup

In order to analyse Subversion, we have prepared an experimental setup which include gathering the data in GitHub and Tigris, perform some queries to join and prepare this information, a statistical analysis about the content in hands and selection of the relevant features for our model. Finally we created our best classification model comparing a range of classification algorithms and parameter refinement. We evaluated the performance with each algorithm tested. In our data we found some commits that clearly have indicated rework and others that not. Our work was then to train the model with the first set and test it with the unlabeled data. Our main task was the classification of commits as rework sources or not, according to the information about the developer's activity.

4.1 Data Gathering

The data mining activity started by collecting the data. Two very helpful tools were used in this job: CVSAnaly and BICHO. These tools are part of the MetricsGrimoire project and fundamental for the OSS research in FLOSSMetrics project [4]. The tools operate similarly: given a project URL they make automatically all the process of crawling the repository and collecting available (and public) data. Then a SLQ relational database was created mapping the relation between the content properties in tables.

With the tools we have collected around 100350 commits, comprehended between July, 2007 and December 2014. There were 33 developers enrolled with the project (committing) in this period, 15 labeled as members. The original data description gathered is listed on Table 1.

[2] Apache Subversion - https://subversion.apache.org/
[3] Subversion Tigris - http://subversion.tigris.org/
[4] FLOSSMetrics Consortium. (2012)

Table 1. Data description - Data originally gathered

Variable	Type	Value	Sample
commit_id	id	integer	189
type	atribute	text	M
file_id	atribute	integer	3
file_name	attribute	text	main.py
file_path	attribute	text	branches/remove-log-addressing/subversion
file_parent_id	attribute	text	4
revision	atribute	integer	100978
date	attribute	date	2014-07-01 04:38:02
code	attribute	text	yes
message	attribute	text	Follow-up to r1639319: Fix file object lifetime.,*
committer_id	attribut	integer	16
committer_name	attribute	text	stefan.fuhrmann
member	attribute	text	no
rework	target	text	yes/no/?

The main item here is a commit, referenced by the commit-id identifier and composed by the id of the developer, its type (member or not), date, revision etc. The field type comes for the type of commit, mapped by CVSAnaly as one of the types: A) ADDITION – file added; D) DELETION – file deleted; M) MODIFIED – file modified and C) COPY - file copied

In this stage we also gathered the information about the issue, namely searching from the commit-id in the issues table and collecting the type of activity. We created a rework column that is used as the label column. When the type of the issue is DEFECT or REOPENED we automatically find the correspondent commit-id responsible for the error and set the rework as true. In the other cases it is set false.

We ran a few statistical analysis on RapidMiner to evaluate the information collected. Figure 1 offers the distribution of the commits between all committers, an interesting behavior that shows how developers contribute. Figure 2 in other hand shows the distribution of the commits by type of commit. We clearly conclude that the action "M" is the most common as it represents code submission.

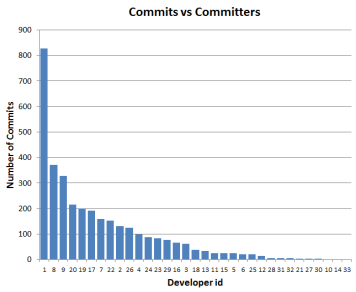


Fig. 1. Distribution of the commits by developers

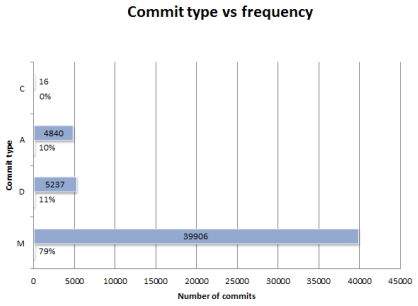


Fig. 2. Commits per type: A) add; D) delete; C) copy; M) modify

We highlight that a great number of commits in the table was not referenced by an issue. In our analysis, only about 35000 of the 100350 commits (near 34%) could be labeled with this approach. This is an important factor for our study, since it means we have a large set of unlabeled data for testing and a correspondent training set (labeled) not so big.

4.2 Data preparation

Using the RapidMiner [5] framework, the next step in the process was conduced. The statistical data analysis implied that not all information contained in the tables was very relevant to our study, and other needed to be prepared. Figure 3 highlights that revision, date, file-path, file-parent-id, committer-name and message seems to have no strong correlation with the rework cases. Although still in Figure 3 we clearly see that the type of the developer was a good property for this. Thinking about the semantics of the rework, and the statistical analysis, the most important features were selected: type of developer, type of modification and number of files affected by one commit. In fact, this information was not present in our initial data, but is a result of counting the files affected by the same commit.

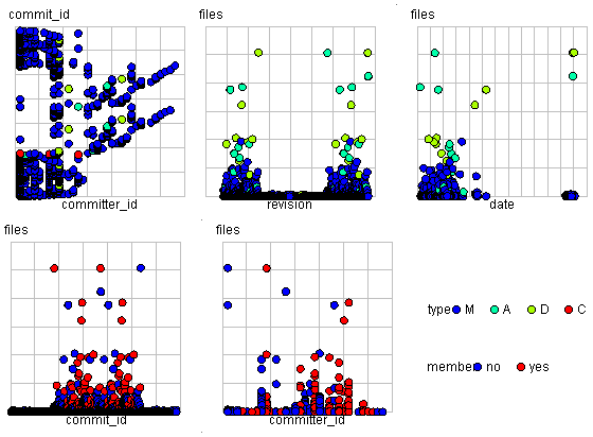


Fig. 3. Statistical visualization of the features

Regarding to data cleansing, we have found some inconsistent commits that were removed. Some had incomplete fields (the tools used do not make all fields mandatory and we encountered many null values that could induce an error in our evaluation), others were identified as **merge** commits (a commit that make

[5] https://rapidminer.com/

the new code available as the official working branch) and were not referenced by activities. There were few outliers identified and we believe that this could be explained by a previous filter in the gathering tools. To solve this situation we have created some basic SQL scripts that filter the undesirable behavior.

Here, our working dataset was consisting in 3311 commits, resulting from the data preparation. We divided it into two subsets: training (containing all labeled samples) consisting of 1490 samples and test (unlabeled) consisting of 1821 samples. We used the training subset for all modelling activities and the test subset for the final evaluation of the model.

4.3 Modeling

In this section, we present the creation of the prediction. More than creating a simple model, we wanted to compare how different algorithms could impact on the results and the knowledge over the data. Using RapidMiner, we've chosen three well-known algorithms for the prediction task: Decision Trees, Naive Bayes and K-NN. These are powerful yet simple mechanisms used in data mining for classification. Moreover, the chosen algorithms needed to be flexible on parameter types and resilient enough for the size and complexity of our dataset. For all three algorithms we have followed the same methodology for improving the model and check performance.

We have started with a simple algorithm model in RapidMiner and applied it to our training set. This implementation gave us a first look at the computed data and a comparative output for improving the performance. The main problem of this approach was that it lacked a training mechanism and could be biased by the distribution of the data. The following approach was improve it by using a 70-30 split mechanism. We sampled the training set using stratified sampling and added a performance validation tool in order to compare results and refine parameters. The final configuration results on different approaches for training, were: Bootstrap, Split-Validation, Cross-Validation and Holdout.

Decision Trees We have first applied the above methodology to the Decision Trees algorithm. We have chosen this first because of the meaning of its results. The algorithm creates a representation of the tree that is the model learned. With Figure 4 we observed the resulting model as the tree output. We could see that our first hypothesis was someway correct as decision weight was heavier on the number of files, membership and type of modification. As we thought, a commit that alters a big number of files have higher probability to introduce rework. Equally, we observed that a commit changing files tends to introduce more rework than a commit removing or adding files, as the last operations were not too frequent as the first.

Following our strategy, we have implemented the other configurations. We have tried different parameter settings and verified that the best setup is accuracy as the parameter criterion and maximal depth 6. Table 2 show the performance evaluation for each configuration. We observed that the results were not

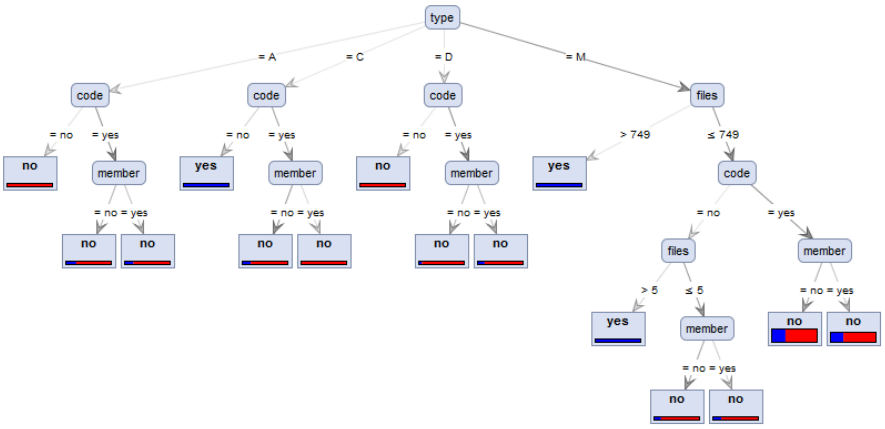


Fig. 4. Decision Tree resulting model

so divergent, which could be a sign that our data represents well the big picture of the project repositories or it that this data do not allow great improvement.

Naive Bayes Differently from Decision Trees, Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. As made with Decision Trees, we have create a simple model and then tried different configurations with Naive Bayes. Table 2 shows the performances of the this algorithm, when applied to our training set and compared to the previous.

K-NN Finally, we have tested the K-NN algorithm. The k-Nearest Nearest Neighbors algorithm is a non-parametric method used for classification and regression that consists in searching the belongings of a sample by consulting the weights of it K-nearest neighbors. We used parameter k = 2 with weighted vote and Mixed Euclidean Distance. Back in Table 2 we could observe that this algorithm clearly outperformed the others.

4.4 Performance and Final Model

In data mining, evaluating performance is very important to verify if our predictions are as good as we want to. Performance is the measure of the correct predictions compared to the errors. Our model was based on identifying commits as rework-sources (or not). We have a true positive when we identify a rework-source as such; not identifying it as so causes a false negative.

From the relationship of these values we inferred how good our classification was using several standard metrics. In our work we used the most relevant

Table 2. Performance evaluation for different setups and different algorithms

Strategy	Accuracy (%)	Recall (%)	Precision (%)
Decision Trees			
Simple Model	69.80	0.22	100
Bootstrap	69.73	0.44	100
Split Validation	69.80	0.22	100
Cross Validation	69.80	0.22	100
Holdout	69.87	0.44	100
Naive Bayes			
Simple Model	70.00	0.44	61.11
Bootstrap	70.00	0.68	60.30
Split Validation	70.00	0.55	100
Cross Validation	70.20	1.48	100
Holdout	70.00	1.71	61.49
K-NN			
Simple Model	75.70	5.19	33.89
Bootstrap	78.52	5.48	36.64
Split Validation	75.70	5.19	31.89
Cross Validation	75.70	5.19	37.34
Holdout	73.83	1.41	100

ones: precision, recall, and accuracy. These values depend on the threshold value obtained using a ROC curve. Figure 5 offers the ROC curves for the three algorithms of the setup. The best threshold depends on the algorithm and by our analysis we inferred that for Decision Trees is 0.42, and for the other two algorithms is 0.5.

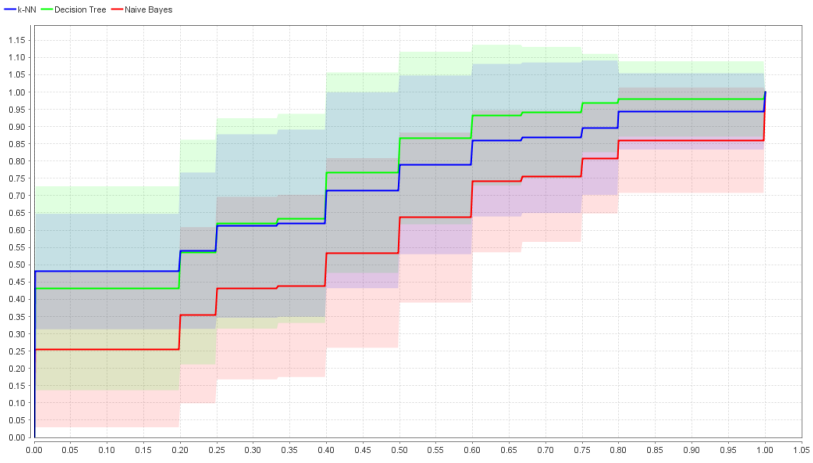


Fig. 5. ROC curve comparison for the threshold

In our problem, accuracy was our main performance metric used to select the best model, our final configuration for the classifier. Accuracy was calculated as seen in Equation 1

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1}$$

and provided a measure to determine the quality of the classification criteria. We have finalized this modeling stage by selecting the K-NN classifier with Bootstraping validation as our final model setup.

5 Empirical Evaluation

We have seen how the final classifier model was created and selected, altogether with the comparison of different algorithms using RapidMiner. But as said in section 4.1, the training subset corresponded to 34% of the total number of commits and the other 66% of the data was called test subset. This was the data we wanted to classify. We wanted to apply our model to this data and check the knowledge that extracted from it. Our ultimate goal was to verify if there were significant developer profiles with different behavior concerning the introduction of rework in OSS code.

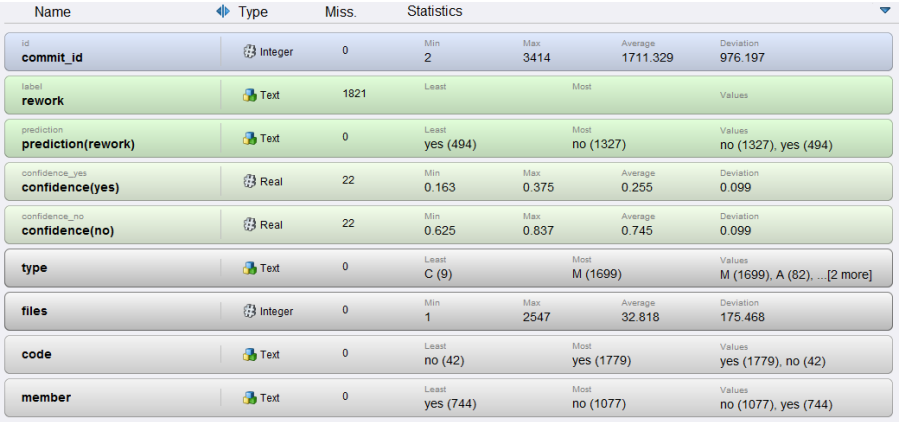
Table 3. Payoff Matrix for the model

	True yes	**True no**	**Class Precision**
Pred. yes	1582	3347	32.10%
Pred. no	1788	4295	70.61%
Class Recall	46.94%	56.20%	

Table 3 shows the training payoff table for the model. The results were obtained applying the classifier to the unlabeled data and condensed in Figure 6. At total, 1821 commits were classified and 494 were predicted to be rework-sources, corresponding to a 27% of the total. Considering that our prediction was realistic, this is a very interesting result. It confirms that rework exists in this environment and seems to achieve the 30%-40% of cost estimated by the industry [8].

Manipulating the data we queried other relevant information. We have discovered that in the total of 494 problematic commits (possible rework sources), 200 (40%) would be introduced by members and 294 (60%) would be in hands of volunteers. From the point of view of the type of rework, 483 (98%) would be in coding activities, facing 11 (2%) in other kinds.

We have found that volunteers introduced more rework in project than members, corroborating the hypothesis that they may not know much about the code specificity. But it is interesting that members of the team introduced a remarkable quantity of rework also. In fact, this is mostly related to the complexity of



Name	Type	Miss.	Statistics			
commit_id	Integer	0	Min: 2	Max: 3414	Average: 1711.329	Deviation: 976.197
rework	Text	1821	Least:	Most:	Values:	
prediction(rework)	Text	0	Least: yes (494)	Most: no (1327)	Values: no (1327), yes (494)	
confidence(yes)	Real	22	Min: 0.163	Max: 0.375	Average: 0.255	Deviation: 0.099
confidence(no)	Real	22	Min: 0.625	Max: 0.837	Average: 0.745	Deviation: 0.099
type	Text	0	Least: C (9)	Most: M (1699)	Values: M (1699), A (82), ...[2 more]	
files	Integer	0	Min: 1	Max: 2547	Average: 32.818	Deviation: 175.468
code	Text	0	Least: no (42)	Most: yes (1779)	Values: yes (1779), no (42)	
member	Text	0	Least: yes (744)	Most: no (1077)	Values: no (1077), yes (744)	

Fig. 6. Results of prediction on unlabeled data

the activity, as members usually have commits with high number of file modifications. Volunteers seemed to choose simpler issues, in which they have lower probabilities to introduce rework. 20 of the 33 developers (60%) introduced rework, and the distribution of the problematic commits is normalized.

The results have shown a clearly coexistence of two profiles of developers by analysing the classification of the commits. We attribute this profile distribution into two distinct roles of developers: (i) regular developers which, members or not, have probability of introducing bugs or other rework in the project and (ii) leader developers, expert in programming techniques with a very good knowledge about the code.

6 Conclusions and Future Work

Rework is a common problem in Software Engineering and some questions about it are open opportunities of research. This work presented a novel approach for the rework problem. Based on data mining we have created a model that can classify developers' actions and predict if they will lead to rework. Moreover, our model could give some precious information about developers' behavior and associate profiles.

The model was developed following a data mining methodology and in the process we have illustrated all the stages from data preparation to performance comparison. Three different algorithms were compared, named Decision Trees, Naive Bayes and K-NN. The results show that the Decision Tree algorithm is a very good way to understand the process, but the best algorithm in our case was K-NN. It outperformed the others in terms of accuracy, recall and precision.

Although volunteers introduced more rework than members confirming our expectations, results have shown that members also presented this behavior. We

have discovered that volunteers usually chose to contribute with more punctual activities, with low impact on the source code. Regarding the classification of commits, we have observed that the most important features for the classification were the number of files affected, the type of action and the role of the developer. Other features such the management of the rework inside project's versions and the impact of the contribution of volunteers in many projects at the same time are beyond the scope of our model.

We consider this work a first attempt to tackle the problem of calculating rework on OSS and to analyse the characteristics and impact of developers in open development. We expect to expand this study in three main ways 1) creating a more general model that applies to all open source repositories and have a better tunneled parameters; 2) managing to develop new algorithms and combined strategies for multi-label classification; 3) integrating this approach in an online tool that could help developers at the moment of merging code.

7 Acknowledgements

The work of Thiago Rúbio has been funded through a IBRASIL Grant. IBRASIL is a Full Doctorate programme selected under Erasmus Mundus, Action 2 – STRAND 1, Lot 16 and coordinated by University of Lille. The second author thanks Universidade do Estado de Mato Grosso (UNEMAT - Brazil), for the support given.

References

1. Gaff, B.M., Ploussios, G.J.: Open source software. Computer **45**(6) (2012) 9–11
2. Boehm, B.: Software risk management. Springer (1989)
3. Zhao, X., Osterweil, L.J.: An approach to modeling and supporting the rework process in refactoring. In: Software and System Process (ICSSP), 2012 International Conference on, IEEE (2012) 110–119
4. Von Krogh, G., Haefliger, S., Spaeth, S., Wallin, M.W.: Carrots and rainbows: Motivation and social practice in open source software development. Mis Quarterly **36**(2) (2012) 649–676
5. Sen, R., Singh, S.S., Borle, S.: Open source software success: Measures and analysis. Decision Support Systems **52**(2) (2012) 364–372
6. Zhu, K.X., Zhou, Z.Z.: Research note-lock-in strategy in software competition: Open-source software vs. proprietary software. Information Systems Research **23**(2) (2012) 536–545
7. Raja, U., Tretter, M.J.: Defining and evaluating a measure of open source project survivability. Software Engineering, IEEE Transactions on **38**(1) (2012) 163–174
8. Zahra, S., Nazir, A., Khalid, A., Raana, A., Majeed, M.N.: Performing inquisitive study of pm traits desirable for project progress. International Journal of Modern Education and Computer Science (IJMECS) **6**(2) (2014) 41
9. Dabbish, L., Stuart, C., Tsay, J., Herbsleb, J.: Social coding in github: transparency and collaboration in an open software repository. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM (2012) 1277–1286
10. Russell, M.A.: Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " O'Reilly Media, Inc." (2013)

SESSION 3

RESEARCH ON PROGRAMMING

Analysis and Evaluation of gesture recognition using LeapMotion
Pedro Leitão

Survey on Frameworks for Distributed Computing
Telmo Morais

Procedural Generation of Maps and Narrative Inclusion for Video Games
João Ulisses, Ricardo Gonçalves and António Coelho

Analysis and Evaluation of Gesture Recognition using LeapMotion

Pedro Miguel Oliveira Leitão

Student of Doctoral Program of Informatic Engineering
Faculty of Engineering, University of Porto
Porto, Portugal
pedromoleitao@gmail.com

Abstract Nowadays are emerging increasingly natural interaction devices, which use human body as a natural way of interacting with applications. To correctly interact with natural interfaces, also named NUI, there is a need to improve the recognition and performance evaluation of different gestures simultaneously in order to identify which configurations between gestures and settings best fit to get a greater efficiency on their recognition. The evaluation was performed based on real gestural attempts with two participants. Finally, the application got a gesture recognition average rate of 86.1% using only the minimum resources provided by the LeapMotion device.

Keywords: Natural User Interface, LeapMotion, Gesture Recognition

1 Introduction

The increasing interest for new interaction paradigms, combined with new emerging technologies, are originating new Natural User Interface (NUI) devices on the market.

Recent devices are aiming to bridge some existing limitations on human-machine interaction through gestures performed with hands, fingers and drawing tools. This kind of devices, such as LeapMotion, which will be described in the next section, are interesting to enthusiastic public and all community due to its simplicity, efficiency and numerous areas where it can be applied. As gestures have distinct characteristics, such as the way and even how fast movements are performed.

On work produced by Sharad Vikram, Lei Li and Stuart Russel [1] authors present an interface of online recognition, of gestures using NUI interaction, performing a very precise interpretation which makes it ideal to drawn words in real-time.

Over time are appearing different techniques associated to Human-Computer-Interaction, based on computational real-time vision, as described by Eshed Ohn-Bar and Mohan Manubhai [2]. They propose a robust system based on natural interaction, to recognize signs in real-time inside a vehicle. These devices used inside vehicles enable a decrease of driver visual charge, driving mistakes and have a high level of adaptation and usability by users.

This paper is organized in the following way: in sections 2 and 3 some NUI devices will be presented. They are equivalent to LeapMotion, but their functioning rely on different available technologies. In section 4 all work done and technical details, about it, will be described. In section 5 all evaluations related to this project will be presented. Finally, on sections 6 and 7, this paper will be concluded and presented some perspectives about future work are discussed.

2 LeapMotion

LeapMotion company was founded in 2010 by Michael Buckwald and David Holz. Working at nearly 300 frames per second, this device (Figure 1) has the capacity to collect hands movements simultaneously, with precision higher than 0.01mm [3].

Figure 1 – LeapMotion device [7].

2.1 Hardware

This device consists of 2 Infrared (IR) monochrome cameras and 3 IR LEDs. As Microsoft Kinect LED, they project a pattern of points along the area which will be captured by IR cameras, collect all data and transferring it to the software layer for analysis purposes. This software uses received data and generate a representation of mapped data, in a three dimensional space to compare with bi-dimensional frames and submit this bi-dimensional images, to an algorithm of edges detection [4].

In order to optimize interaction performance, process of transfer information over USB cable is submitted to a compression process in order to remove background light and unduly added noise. After that, three-dimensional data collected by IR sensors will be analysed and reconstructed for representation. Finally, to transfer information to LeapMotion API, an algorithm of tracking,

looking for information about hands, fingers and tools representing them on three-dimensional and transferring this data into API layer responsible for communicate with high level software [5].

The biggest challenge of LeapMotion team still be the residual latency. This problem is related with captured images, at a moment that will suffer delay and will only show these images some milliseconds after the movement has completed.

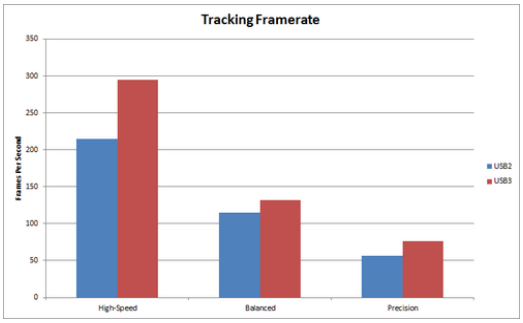
The better way to minimize this problem and improve processor response time should consider the follow settings [6]:

- Use USB3.0 cable to higher transfer rate;
- Use monitor with short response time;
- Initiate LeapMotion in 'High-Speed' mode.

2.2 Technical Details

The device LeapMotion, has 3 different operating modes presented on graphic (Graphic 1): 'High-Speed', 'Balanced' e 'Precision':

- The method 'High-Speed' is suitable for scanning fast movements. This mode increases a resolution of IR sensors, making them quadruple the number of captured images per second of data collection of movements.
- Next mode is 'Precision' mode, with this it is possible to decrease frame rate to 40% doing it lower than the normal value, without need to decrease resolution. It is ideal for capture of small movement variations.
- Finally, with 'Balanced' mode, it tries to reconcile recognition features as quickly and accurately, adapting referred modes in just one, balancing the resolution bandwidth and computational charge with the value of the frame rate.



Graphic 1 - Frame rates and operationing modes of LeapMotion using USB cables 2.0 and 3.0 [5].

One feature which should be considered is about data acquisition and what kind of cable used and processing capacity of used machine.

With USB cable 3.0, it is possible to transfer a higher data frame rate comparing with it is predecessor USB2.0. However, with this cable 3.0 which ensure higher

transfer of data needs a high computer capacity to be able to process all received data, and use this higher amount of information to reduce of movements delay.

2.3 Planes of Interaction

To detect touch in deepness over a tri-dimensional interaction zone, 3 zones of detection are automatically defined (Figure 2). These 3 plans, represented with different colors, are bounding different zones of interaction.

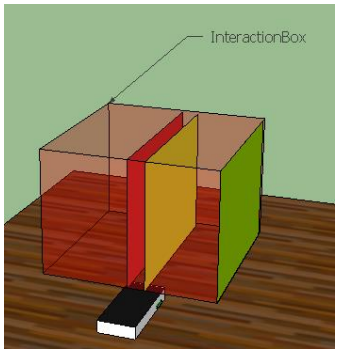


Figure 2 – Interaction zones to detecting touch on the areas.

With a selected object to interact with those zones and going forward over the interaction box, our object will find the following plans: 'None'(green), 'Hovering'(yellow) and 'Touching'(red) on (Figure 2). Starting at user position and entering into interaction box bounds, the first interaction zone is called 'None' bounded between 130 and 280 mm and is used to navigate over this bi-dimensional zone. Interaction zone 'Hovering' is between 70 and 130 mm is here where object navigator would be positioned over bi-dimensional coordinate which will be clicked. Last interaction zone is called 'Touching' settled between -100 and 70 mm and is used to seal action of touch in a coordinate selected on previous zone 'Hovering'.

2.4 Software

LeapMotion SDK is supported by Windows, Macintosh and Linux operating systems. First version of this SDK, offers applications to test as well as functions to calibrate device to required settings [8]. The second version of this SDK already exists on experimental version [9] which includes as main improvements the capability to represent all structure of a hands articulations simultaneously, more

robustness to interference originated by sunlight and the implementation of new gestures such as pinch and grab objects [10].

3 Application Development

The scope of this project is intended to develop an evaluation application developed on .Net Framework 4, named Leap Tester composed by two interfaces to evaluate individual and a group of gestures, giving two types of analysis and evaluations. On first stage was started an individual analysis of each gesture type and some of their parameters. On second stage, was focus on a general evaluation of all gestures simultaneously, using better results of previous stage in order to get better results.

For each move, were implemented functions which instantiate device commands which could be integrated with different applications to associate this gesture command with different kinds of functions. We selected four types of gestures available with LeapMotion library, which use only one hand, right or left and create two new gestures one of them use two hands and the other with one hand right or left. The last ones were created to measure the recognition accuracy of gestures with greater movement amplitudes. Each function was associated with a device library gesture distinct type. So CLEAN move instance SWIPE command, CLICK DISPLAY is associated with SCREENTAP command, CIRCULAR move to both sides, right and left, are associated to CIRCLE command, at last CLICK function are related to KEYTAP command. The new type gestures implemented were APPROACH TWO HANDS and FIVE FINGERS.

3.1 Gestures Implementation

Of all recognizable signs embedded on LeapMotion library were modified the nature of SWIPE and CIRCULAR gestures in order to check their performance using different movement configurations (Figure 3).

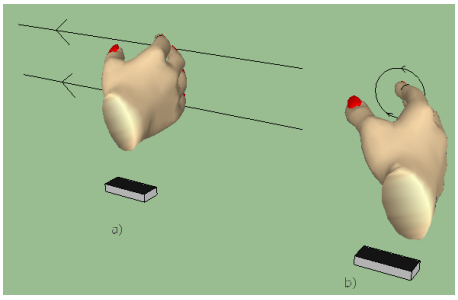


Figure 3 - a) SWIPE move performed horizontally from right to left. b) CIRCULAR move in both directions.

The SWIPE movement could be recognized as vertical or horizontal sliding move, using index finger of both hands. So to restrict this movement to only accept horizontal sweeps performed from right to left side. To bound this horizontal move absolute value of bi-dimensional vector generated by sweep move (1) on X axis should be bigger than vertically on Y axis. About movement direction, if value of D_x was bigger than 0 so, sign is produced on clockwise, otherwise if value is smaller than 0 movement is counter-clockwise.

$$H = |D_x| > |D_y| \quad (1)$$

About CIRCULAR movement, this move allows to do circular moves using both index fingers and production an circle with minimum radius of 5 mm and minimum arc length of 1.5 * π radians. To identify the way the circular movement was produced, calculating the angle of normal vector resultant of this move (2). In case of the angle value being less than 90º, movement was performed clockwise; otherwise it was performed counter-clockwise.

$$\frac{D \cdot C}{\|D\| \, \|C\|} \leq \frac{\pi}{2} \quad (2)$$

Beyond the referred modifications, new two gestures were implemented, the APPROACH TWO HANDS and FIVE FINGERS.

While first implementation of APPROACH TWO HANDS was configured to bound this gesture to be recognized when palm of hands being approached, at a distance less than 4 cm and vertically between them.

The gesture FIVE FINGERS is acknowledged through approximation of two hands of sensor in order to detect its fingers. As device don't know when should start to detect hand fingers, so when a hand is detected inside the interaction box, is performed immediately the function of move detection. This becomes inappropriate when we want the gesture to be submitted to another validation using another method. Based on this adversity, a function was developed to allow user to insert the hand inside the box. After inserting the hand a countdown variable is started performing the recognition of five fingers after elapsed time.

3.2 User Interface

Developed application have two different kind of interfaces to allow participants to make their experiments. This application is split into two groups (Figure 4): evaluation interface with gestures parameterization and in other hand, the interface of test generation with all gestures simultaneously.

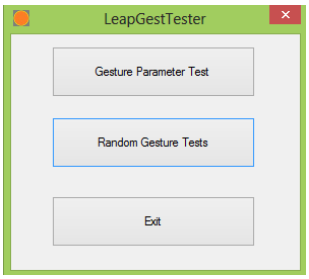


Figure 4 – Selection menu of evaluator.

On first group (Figure 5), user can select any gesture and change their parameters, generating values between a minimum and a maximum values defined by him. When each gesture is recognized, user needs to remove his hand from LeapMotion interaction area to avoid repeated moves and click over 'Next Gesture' button to get another attempt. This procedure is repeated about 50 trials, where user should answer correctly to all requested gestures.

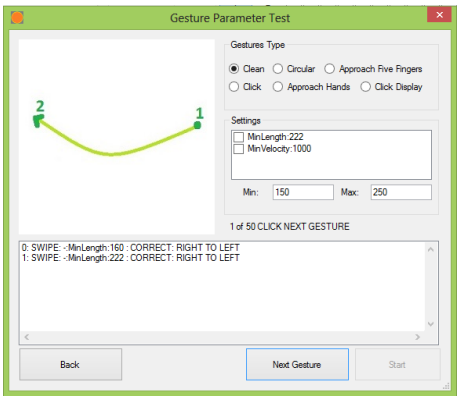


Figure 5 –Interface of evaluation by type.

Then, after finishing all individual sequences of assessment of each gesture and after analyze obtained results. The second group, (Figure 6), is about implementation of retrieved values with better results of previous experience and use these values to initiate a new and last stage of tests to evaluate all gesture types generated randomly.

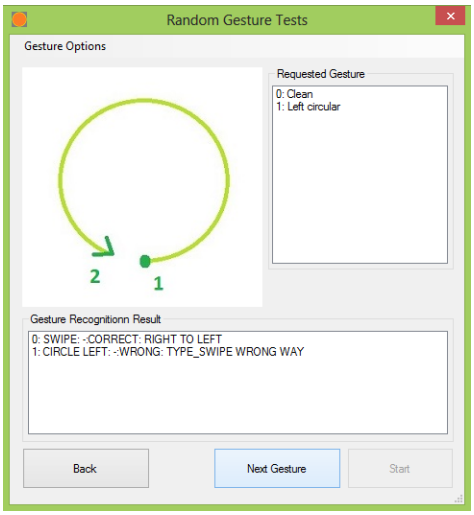


Figure 6 - Interface of general evaluation.

In both experiment interfaces, all acquired results of attempts are saved in text files to be used for further analysis.

4 Validations

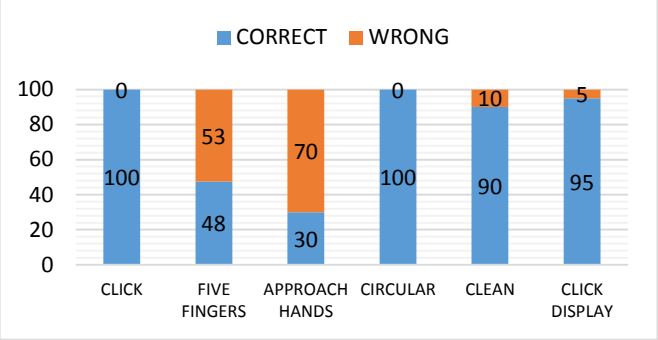
In order to verify the feasibility of the developed application, we proceeded to an informal evaluation with 2 participants with 20 and 27 years old which each of them did 200 attempts, using LeapMotion with USB cable 2.0 at 'Balanced' mode. The reason why we have chosen these settings to perform our experiments over worst conditions in order to optimize the latency problem of our application and under bad conditions try to achieve the best possible results of gestures recognition.

All evaluations, were performed using a machine equipped with a CPU Intel Core i7-4700MQ working at 2.40 Ghz with 4GB RAM DDR3, using USB cable 2.0 and working on operating mode 'Balanced' where it is frame rate was between 110 and 120 fps.

During the implementation of new gestures, and along of experimental stage of gesture FIVE FINGERS, this move was detected and instantly recognizing all fingers when user put his hand over device, entering on interaction box and crossing plans. To improve this situation and add another verification in order to give time to change his command or retreat his action. It was decided to give freedom of movement before recognizing five fingers, and allow the user to move his hand inside the interaction box. Only after a period of time of 5 seconds recognition process of fingers is launched.

In first evaluation stage, was applied general evaluation tester, and generated randomly 600 attempts. All gestures generated were defined to inhibit the remaining

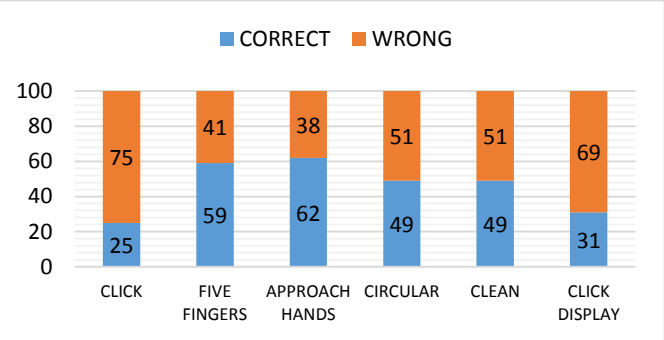
5 types of gestures. Every time, one gesture was asked all others would be disregarded. This test methodology got an average of correct gesture identifications of 77.1% (Graphic 1st).



Graphic 1st – Results of 1º evaluation of all gesture types.

During second evaluation was applied general evaluation tester as on previous stage. However, was removed the restriction about the recognition of one type of gestures in each test and enabled the ability to recognize all gesture types at same time. Now all gestures: CLICK, FIVE FINGERS, APPROACH HANDS, CIRCULAR, CLEAN and CLICK DISPLAY could be recognized on any attempt.

So, each gesture should be much distinct as possible to improve correctly identification. At the moment in which all gestures are accepted, makes CLEAN gesture wrongly recognized, because this move is applied in many others different moves, misunderstandings errors of recognition. Through elaboration of a set of more than 600 attempts was possible to verify the number of correct answers went down significantly to 45.8% (Graphic 2nd).

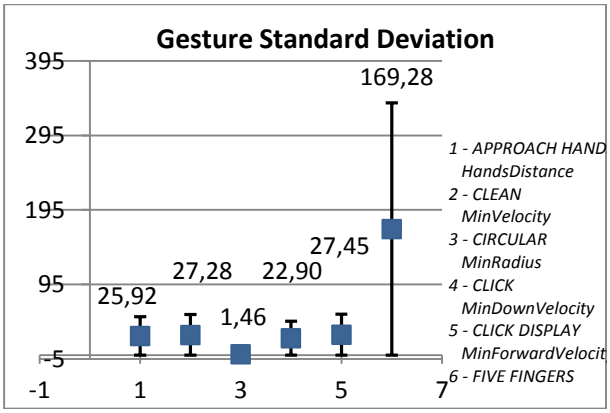


Graphic 2nd - Results of 2º evaluation of all gesture types.

Among second evaluation and the last, was considered individually each type of gesture changing some parameters of them and testing each one separately. For APPROACH TWO HANDS gesture, was changed distance between hands through parameter *HandsDistance;* generating randomly values between 100 and 200 mm

and replacing them in order to find best value which was 142 mm. On CLEAN gesture, was changed swipe minimum velocity parameter *MinVelocity,* which changed between 150 and 250 mm/s, getting 209.2 mm/s as best result. Next gestures were two circular directions of CIRCULE which was randomly, changed the radius *MinRadius* of the circle move using finger replacing by values between 5 mm and 10 mm and getting 6.7 mm as better result. Follow gesture CLICK DISPLAY, was replaced value of minimum velocity of finger getting in on interaction box *MinForwardVelocity*, the speed variation was between 50 and 150 mm/s getting an optimal value of 93 mm/s. Next gesture was the CLICK, and parameter changed on him was click minimum velocity, replacing with values between 50 and 100 mm/s and getting 87.2 mm/s as best result. Finally, on detection of 5 FINGERS was replaced parameter *DetectHandFingers* with values between 1000 and 1500 ms having 1290.8 ms as optimal value.

On next graph (Graphic 2), are presented all types of gestures as well as dispersion of values collected based in the average of each gesture. From all gestures CIRCULAR as the one who had lower variability of values, on other hand FIVE FINGERS was the one who had more different correct values along the recognitions.



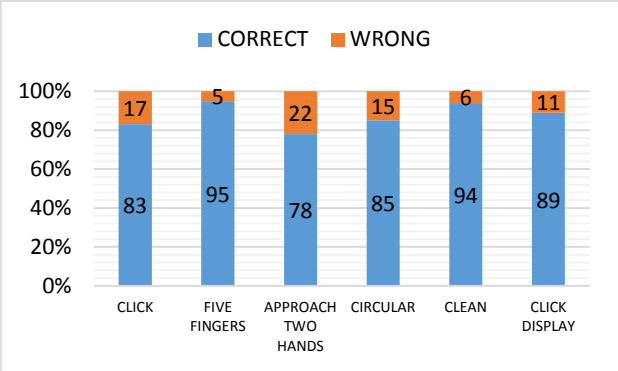
Graphic 2 - Standard deviations of the values of each parameter applied to each gesture.

At a final evaluation stage, was planned a strategy to improve recognition, enabling in each request gesture could be recognize another one. Combination of gestures as shown on table below (Table 1) for each required gesture, application can only identify the requested sign and another one with the exception of FIVE FINGERS and APPROACH TWO HANDS in these last, is possible to recognize all types of gestures. All these combinations are grouped based on gestures differences joining those with more differences between themselves.

Request	Gestures combinations			
CLEAN	CLEAN	CIRCULAR		
CIRCULAR	CIRCULAR	CLICK		
CLICK	CLICK	CLEAN		
CLICK DUSPLAY	CLICK DUSPLAY	CLICK		
FIVE FINGERS	CLEAN	CIRCULAR	CLICK	CLICK DISPLAY
APPROACH TWO HANDS	CLEAN	CIRCULAR	CLICK	CLICK DISPLAY

Table 1 – Combination of allowed gestures on 3º evaluation.

Through the collected values from individual experiments, last evaluating stage were 86.1% better using this method than others prior methods applied on previous evaluations. This method allows to reduce problems about affinities between signs and grouping each of them in order to improve signs recognitions as shown on graphic below (Graphic 3rd). APPROACH TWO HANDS and FIVE FINGERS are gestures which are very heterogeneous, comparing features between them and with all others gestures, so don't was necessary restrict them. Throughout the reviews, a common error was the detection of CLEAN sign who create interferences because the sliding movement some time is associated with a begin of an CLICK gesture when finger starts the approach.



Graphic 3rd - Results of 3º evaluation of all gesture types.

5 Conclusion and Future Work

The study of different configuration features applied to gestures allow us to identify a set of parameters which can perform a low recognition conflict between them and a good results with different features between their recognition working together. At the moment with experiments performed to people with different ages was

possible to determine the feasibility of this method applied with adults. Although the results indicate the young public shows more difficult along the familiarization process with plans, comparing with the older participant. However, the younger participant shows to have on other hand greater adaptation capacity, gradually improving his coordination capacities.

Along the evaluations, acquired results allow us to find some of better values to improve recognition process of gestures in order to get better results on sign recognitions.

As perspective of future work, we would like to do more experiments in order to test more deeply our method. Furthermore add new gestures using new techniques and approaches in order to get better efficiency results and bridging ambiguity of signs. Another possibility would be the upgrade to evaluate new version of LeapMotion SDK using a stable release of this version to get the degree of efficiency, flexibility and robustness of this new library.

To conclude, with the improvement achieved on the last experiment, was possible to approach a possible implementation of gestures applied. On 3rd evaluation stage had a very good recognition rate, making these settings a serious candidate to integrate NUI application to draw and recognize characters written in the air [11].

References

1. Vikram S., Li L., Russel S. *Handwriting and Gesturein the Air, Recognition on the Fly*, University of California, Berkeley, 2013.
2. Ohn-Bar E.,Trivedi M. Hand *Gesture Recognition for Automotive Interfaces: A Multimodal Vision-based Approach and Evaluations*. IEEE Transactions on Intelligent Transportation Systems 2014, 2014.
3. LeapMotion. Leap Motion Controller Set To Ship May 13 for Global Pre-Orders, February, 2013.
4. Nguyen T., Schulze J., Kiyokawa K. Sculpting in Real-Time Using a Leap Motion, University of California, San Diego; Osaka University, 2013.
5. Bedikian R., 2014. Understanding Latency: Part 1. Leap Motion Blog. July 2013.
6. Bedikian R., 2014. Understanding Latency: Part 2. Leap Motion Blog. July 2013.
7. Westover B., 2013. Leap Motion Controller review, July 2013.
8. Pinto P., 2013. Leap Motion – Unboxing e Análise, April 2013.
9. Leap Motion Developer Portal, July 2014.
<https://developer.leapmotion.com/>
10. mbuckwald, 2014. V2 Tracking Now in Public Developer Beta, May 2014
<https://community.leapmotion.com/t/v2-tracking-now-in-public-developer-beta/1202>
11. Pedro Leitão; João J. Pereira; António Castro (2012), Interface caligráfica de escrita no ar, 20º EPCG: IPVC.

Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm

Telmo da Silva Morais

Student of Doctoral Program of Informatics Engineering
Faculty of Engineering, University of Porto
Porto, Portugal
Telmo.morais@gmail.com

Abstract
The storage and management of information has always been a challenge for software engineering, new programing approaches had to be found, parallel processing and then distributed computing programing models were developed, and new programing frameworks were developed to assist software developers. This is where Hadoop framework, an open source implementation of MapReduce programing model, that also takes advantage of a distributed file system, takes its lead, but in the meantime, since its presentation, there were evolutions to the MapReduce and new programing models that were introduced by Spark and Storm frameworks, that show promising results.

Keywords: Programing framework, Hadoop, Spark, Storm, distributed computing.

1 Introduction

Through time, size of information kept rising and that immense growth generated the need to change the way this information is processed and managed, as individual processors clock speed evolution slowed, systems evolved to a multi-processor oriented architecture. However there are scenarios, where the data size is too big to be analysed in acceptable time by a single system, and in this cases is where the MapReduce and a distributed file system are able to shine.

Apache Hadoop is a distributed processing infrastructure. It can be used on a single machine, but to take advantage and achieve its full potential, we must scale it to hundreds or thousands of computers, each with several processor cores. It's also designed to efficiently distribute large amounts of work and data across multiple systems.

Apache Spark is a data parallel general-purpose batch-processing engine. Workflows are defined in a similar and reminiscent style of MapReduce, however, is much more capable than traditional Hadoop MapReduce. Apache Spark has its Streaming API project that allows for continuous processing via short interval batches. Similar to Storm, Spark Streaming jobs run until shutdown by the user or encounter an unrecoverable failure.

Apache Storm is a task parallel continuous computational engine. It defines its workflows in Directed Acyclic Graphs (DAG's) called "topologies". These topologies run until shutdown by the user or encountering an unrecoverable failure.

1.1 The big data challenge

Performing computation on big data is quite a big challenge. To work with volumes of data that easily surpass several terabytes in size, requires distributing parts of data to several systems to handle in parallel. By doing it, the probability of failure rises. In a single-system, failure is not something that usually program designers explicitly worry about.[1]

However, in a distributed scenario, partial failures are expected and common, but if the rest of the distributed system is fine, it should be able to recover from the component failure or transient error condition and continue to make progress. Providing such resilience is a major software engineering challenge. [1]
In addition, to these sorts of bugs and challenges, there is also the fact that the compute hardware has finite resources available. The major hardware restrictions include:

- Processor time
- Memory
- Hard drive space
- Network bandwidth

Individual systems usually have few gigabytes of memory. If the input dataset is several terabytes, then this would require a thousand or more machines to hold it in RAM and even then, no single machine would be able to process or address all of the data.

Hard drives are a lot bigger than RAM, and a single machine can currently hold multiple terabytes of information on its hard drives. But generated data of a large-scale computation can easily require more space than what original data had occupied. During this, some of the storage devices employed by the system may get full, and the distributed system will have to send the data to other node, to store the overflow.
Finally, bandwidth is a limited resource. While a pack of nodes directly connected by a gigabit Ethernet generally experience high throughput between them, if all transmit multi-gigabyte, they would saturate the switch's bandwidth. Plus, if the systems were spread across multiple racks, the bandwidth for the data transfer would be more diminished [1].

To achieve a successful large-scale distributed system, the mentioned resources must be efficiently managed. Furthermore, it must allocate some of these resources toward maintaining the system as a whole, while devoting as much time as possible to the actual core computation[1].

Synchronization between multiple systems remains the biggest challenge in distributed system design. If nodes in a distributed system can explicitly communicate with one another, then application designers must be cognizant of risks associated with such communication patterns. Finally, the ability to continue computation in the face of failures becomes more challenging[1].

Big companies like Google, Yahoo, Microsoft have huge clusters of machines and huge datasets to analyse, a framework like Hadoop helps the developers use the cluster without expertise in distributed computing, and taking advantage of Hadoop Distributed File System.[2]

2 State of the Art

This section will begin to explain what is Apache's Hadoop Framework and how it works, also a short presentation of other Apache alternative frameworks, namely Spark and Storm.

2.1 The Hadoop Approach

Hadoop is designed to efficiently process large volumes of information by connecting many commodity computers together to work in parallel. One hypothetic 1000-CPU machine would cost a very large amount of money, far more than 1000 single-CPU or 250 quad-core machines. Hadoop will tie these smaller and more reasonably priced machines together into a single cost-effective compute cluster.[1]
Apache Hadoop has two pillars:

- YARN - Yet Another Resource Negotiator (YARN) assigns CPU, memory, and storage to applications running on a Hadoop cluster. The first generation of Hadoop could only run MapReduce applications. YARN enables other application frameworks (like Spark) to run on Hadoop as well, which opens up a wide set of possibilities.[3]
- HDFS - Hadoop Distributed File System (HDFS) is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system.[3]

MapReduce
Hadoop is modelled after Google MapReduce. To store and process huge amounts of data, we typically need several machines in some cluster configuration.

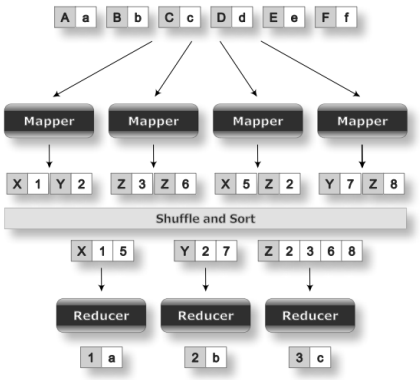


Fig. 1. - MapReduce flow

A distributed file system (HDFS for Hadoop) uses space across a cluster to store data, so that it appears to be in a contiguous volume and provides redundancy to prevent data loss. The distributed file system also allows data collectors to dump data into HDFS, so that it is already prime for use with MapReduce. Then the Software Engineer writes a Hadoop MapReduce job [4].

Hadoop job consists of two main steps, a map step and a reduce step. There may be, optionally, other steps before the map phase or between the map and reduce phases. The map step reads in a bunch of data, does something to it, and emits a series of key-value pairs. One can think of the map phase as a partitioner. In text mining, the map phase is where most parsing and cleaning is performed. The output of the mappers is sorted and then fed into a series of reducers. The reduce step takes the key value pairs and computes some aggregate (reduced) set of data, i.e. sum, average, etc [4].

The trivial word count exercise starts with a map phase, where text is parsed and a key-value pair is emitted: a word, followed by the number "1" indicating that the key-value pair represents 1 instance of the word. The user might also emit something to coerce Hadoop into passing data into different reducers. The words and 1s are sorted and passed to the reducers. The reducers take like key-value pairs and compute the number of times the word appears in the original input.[5]

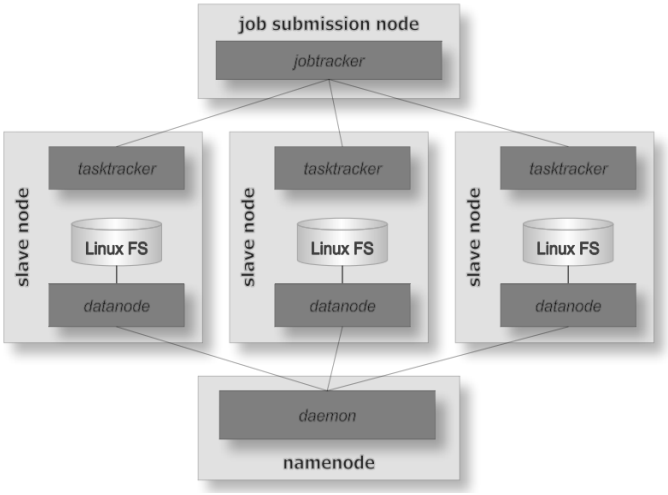


Fig. 2. - Hadoop workflow[2]

2.2 SPARK framework

Apache Spark is an in-memory distributed data analysis platform, primarily targeted at speeding up batch analysis jobs, iterative machine learning jobs, interactive query and graph processing. One of Spark's primary distinctions is its use of RDDs or Resilient Distributed Datasets. RDDs are great for pipelining parallel operators for computation and are, by definition, immutable, which allows Spark a unique form of fault tolerance based on lineage information. If you are interested in, for example, executing a Hadoop MapReduce job much faster, Spark is a great option (although memory requirements must be considered) [19].

It provides high-level APIs in Java, Scala and Python, and an optimized engine that supports general execution graphs.

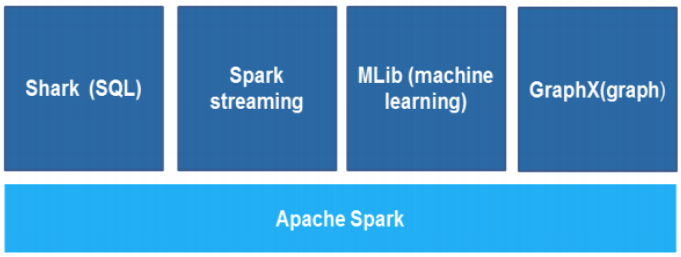


Fig. 3. - Spark Framework

It also supports a rich set of higher-level tools, including Shark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming [6], helping the development of parallel applications.

The main goal of Spark is to work with distributed collections, as you would with local ones. It relays on a resilient distributed datasets (RDDs), that is a immutable collections of objects spread across a cluster, built through parallel transformations (map, filter, etc), automatically rebuilt on failure controllable persistence (e.g. caching in RAM) for reuse, shared variables that can be used in parallel operations [6].

Resilient Distributed Datasets (RDDs)

Spark's main abstraction is resilient distributed datasets (RDDs), which are immutable, partitioned collections that can be created through various data-parallel operators.

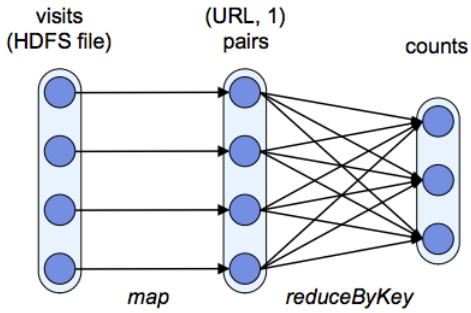


Fig. 4. - Lineage graph for the RDDs in our Spark example.[7]

Each RDD is either a collection stored in an external storage system, such as a file in HDFS, or a derived dataset created by applying operators to other RDDs. For example, given an RDD of (visitID, URL) pairs for visits to a website, we might compute an RDD of (URL, count) pairs by applying a map operator to turn each event into a (URL, 1) pair, and afterward a reduce to add the counts by URL.[7]
Spark provides three options for persist RDDs:

1. In-memory storage as deserialized Java Objects (fastest, JVM can access RDD natively) [2].
2. In-memory storage as serialized data (space limited) [2].
3. On-disk storage (RDD too large to keep in memory, and costly to recomputed) [2].

Spark streaming

The key idea behind the model is to treat streaming computations, as a series of deterministic batch computations, on small time intervals. The input data received during each interval is stored, reliably across the cluster, to form an input dataset for that interval. Once the time interval completes, this dataset is processed via deterministic parallel operations, such as *map*, *reduce* and *groupBy*, to produce new datasets repre-

senting program outputs or intermediate state. It stores these results in resilient distributed datasets (RDDs)[8].

Apache Spark does not itself require Hadoop to operate. However, its data parallel paradigm requires a shared file system for optimal use of stable data. The stable source can be S3, NFS, or, more typically, HDFS) [9].

2.3 Storm Framework

Apache Storm is, a free and open source distributed real-time computation system, focused on stream processing or what some call complex event processing. Storm implements a fault tolerant method for performing a computation or pipelining multiple computations on an event, as it flows into a system. One might use Storm to transform unstructured data, as it flows into a system into a desired format)[9].

Apache Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Storm has many use cases: real-time analytics, online machine learning, continuous computation, distributed RPC, ETL and more. It's scalable, fault-tolerant, guarantees your data will be processed, is easy to set up and operate [9].

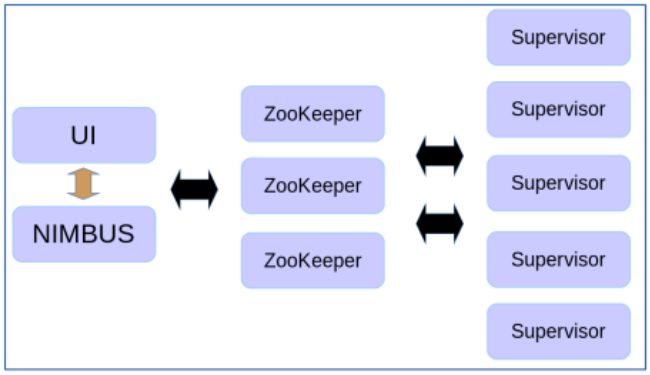


Fig. 5. - Storm Framework system architecture

System architecture:

- Nimbus: Like JobTracker in Hadoop
- Supervisor: Manage workers
- Zookeeper: Store meta data
- UI: Web-UI

A Storm cluster is superficially similar to a Hadoop cluster. Whereas on Hadoop you run "MapReduce jobs", on Storm you run "topologies". "Jobs" and "topologies" themselves are very different; one key difference is that a MapReduce job eventually finishes, while a topology processes messages forever (or until you kill it).

There are two kinds of nodes on a Storm cluster: the master node and the worker nodes. The master node runs a daemon called "Nimbus" that is similar to Hadoop "JobTracker". Nimbus is responsible for distributing code around the cluster, assigning tasks to machines, and monitoring for failures [9].

Each worker node runs a daemon called the "Supervisor". The supervisor listens for work assigned to its machine, starts and stops worker processes, as necessary based on what Nimbus has assigned to it. Each worker process executes a subset of a topology. A running topology consists of many worker processes, spread across many machines [9].

Storm does not natively run on top of typical Hadoop clusters, it uses Apache ZooKeeper and its own master/minion worker processes to coordinate topologies, master and worker state, and the message guarantee semantics [9].

Having said that, both Yahoo! and Hortonworks are working on providing libraries for running Storm topologies on top of Hadoop 2.x YARN clusters.

Regardless, Storm can certainly still consume files from HDFS and/or write files to HDFS[18][21].

3 Discussion

Spark is one of the newest players in the MapReduce field. Its purpose is to make data analytics fast to write, and fast to run. Unlike many MapReduce systems (Hadoop inclusive), Spark allows in-memory querying of data (even distributed across machines) rather than using disk I/O. It's no surprise that Spark out-performs Hadoop on many iterative algorithms. Spark is implemented in Scala, a functional object-oriented language that runs on top of the JVM. Similar to other languages like Python and Ruby, Scala has an interactive prompt that users can use to query big data straight from the Scala interpreter, making it a good choice in some scenarios. However, it does not support a distributed file system on its own, it depends on Hadoop, if a HDFS is required.

The Storm framework is referred as being the Hadoop of Real-time Processing. Hadoop is a batch-processing system, this means, give it a big set of static data and it will do something with it. Storm is real-time, it processes data in parallel as it streams. Therefore, Storm is more a complement to Hadoop rather than a real replacement, as Storm fails when it comes to process large persistent data, as its focus is to be able to process a large number of streams of data (in real time computation), while Hadoop focus is on large amount of persistent data (batch processing).

3.1 Frameworks features summary

In the beginning of this survey, I did not know what I would find on programing frameworks for distributed computing. Therefore, after this review, summarizing the main features and benefits of each of the evaluated frameworks, may serve as contribution to an appropriated and better-informed selection of the framework, that aims

the deployment of a new distributed computing platform, or for those considering to improve an already existing one.

	Storm	Spark Streaming
Processing Model	Record at a time	Mini batches
Latency	Sub second	Few seconds
Fault tolerant – every record processed	At least one (may be duplicates)	Exactly one
Batch framework integration	Not available	Spark
Supported languages	Storm was designed from the ground up to be usable with any programming language [9].	Python, Scala, Java

Table 1. Storm vs. Spark Streaming

Based on my research, the comparison must be made based on use cases oriented view, as the frameworks end up being more complementary than competitive among each other. One thing was made clear, in all references, it does not matter if you choose Hadoop, Spark or Storm, having the HDFS is an advantage, because it solves many of storage problems associated with big data computing. So Hadoop is kind of "mandatory", if you need HDFS benefits.

For Spark, its best use cases, are iterative Machine Learning algorithms and Interactive analytics. Furthermore, Spark plus Hadoop is always better than only Hadoop, except when the work dataset size exceeds the individual node RAM size, so in a way it depends on the available infrastructure or required work dataset size.

Storm is a good choice if you need sub-second latency and no data loss. Spark Streaming is better if you need stateful computation, with the guarantee that each event is processed exactly once. Spark Streaming programming logic may also be easier because it's similar to batch programming, in that way, you are working with batches (albeit very small ones)[19].

One key difference between these two technologies is that Spark performs Data-Parallel computations while Storm performs Task-Parallel computations.

4 Conclusion

After this analysis it is possible to realize that the Hadoop framework will stay around for a while, and for a good reason. Even knowing that MapReduce cannot solve every problem, it is still a good choice for research, experimentation, and everyday data manipulation. One of the other frameworks abovementioned, may be better if the advantages of HDFS are not necessarily imperative, or if the use cases are compatible with the framework capabilities, and consequently able to take advantage of its benefits.

In overall, the most suitable platform must always take into account the scenario to witch the system is most focussed.

It's important to acknowledge that the newer Hadoop versions based on YARN, allow Spark to run on top of Hadoop, and there is on-going work to achieve the same with Storm.

It remains to be seen how successful those implementations are, and also how they compare to its native counterparts versions Spark and Storm, since these aspects weren't approached by this survey.

Acknowledgements

The author takes here the chance to say thanks to all the reviewers anonymous that helped whit their revision to improve the resulting quality of this paper.

References

1. Yahoo! Hadoop Tutorial. https://developer.yahoo.com/hadoop/tutorial/. Accessed 20 Dec 2014.
2. Aridhi S (2014) Frameworks for Distributed Computing Sabeur Aridhi.
3. What is Hadoop. http://www-01.ibm.com/software/data/infosphere/hadoop/. Accessed 22 Dec 2014.
4. Rosario R (2011) No Title. http://www.bytemining.com/2011/08/hadoop-fatigue-alternatives-to-hadoop/. Accessed 15 Dec 2014.
5. Welcome to Apache™ Hadoop®! http://hadoop.apache.org/. Accessed 20 Dec 2014.
6. Apache Spark. https://spark.apache.org/. Accessed 26 Dec 2014.
7. Xin R, Rosen J, Zaharia M (2013) Shark: SQL and rich analytics at scale.
8. Zaharia M, Das T, Li H, et al. (2012) Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters. Proc. 4th Edition.
9. Apache Storm. https://storm.apache.org/. Accessed 27 Dec 2014.
10. Xuhui Liu; Jizhong Han; Yunqin Zhong; Chengde Han; Xubin He, Implementing WebGIS on Hadoop: A case study of improving small file I/O performance on HDFS, Cluster Computing and Workshops, 2009. CLUSTER '09. IEEE International Conference on , vol., no., pp.1,8, Aug. 31 2009-Sept. 4 2009.
11. L. Jiang, B. Li, M. Song, THE optimization of HDFS based on small files, In 3rd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT2010), Beijing, 2010. pp. 912-915.
12. G. Mackey, S. Sehrish, J. Wang, Improving metadata management for small files in HDFS, In 2009 IEEE International Conference on Cluster Computing and Workshops (CLUSTER'09), New Orleans,Sept, 2009, pp.1-4.
13. Jiong Xie; Shu Yin; Xiaojun Ruan; Zhiyang Ding; Yun Tian; Majors, J.; Manzanares, A.; Xiao Qin, Improving MapReduce performance through data placement in heterogeneous Hadoop clusters, Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on , vol., no., pp.1,9, 19-23 April 2010.

14. Thanh, T.D.; Mohan, S.; Eunmi Choi; SangBum Kim; Pilsung Kim, A Taxonomy and Survey on Distributed File Systems, Networked Computing and Advanced Information Management, 2008. NCM '08. Fourth International Conference on, vol.1, no., pp.144,149, 2-4 Sept. 2008.
15. S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In SOSP '03: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, pages 29–43, New York, NY, USA, 2003. ACM.
16. J. M. Hellerstein, M. Stonebraker, and J. Hamilton. Architecture of a database system. Foundations and Trends in Databases, 1(2): 141–259,2007.
17. Apache Storm vs. Apache Spark. http://www.zdatainc.com/2014/09/apache-storm-apache-spark/. Accessed 20 Dec 2014.
18. Storm vs. Spark Streaming: Side-by-side comparison. http://xinhstechblog.blogspot.pt/2014/06/storm-vs-spark-streaming-side-by-side.html. Accessed 20 Dec 2014.
19. How to run Storm on Apache Mesos. https://mesosphere.com/docs/tutorials/run-storm-on-mesos/. Accessed 20 Dec 2014.
20. Storm on YARN Install on HDP2 Cluster. http://hortonworks.com/kb/storm-on-yarn-install-on-hdp2-beta-cluster/. Accessed 20 Dec 2014.

Procedural Generation of Maps and Narrative Inclusion for Video Games

João Ulisses[1], Ricardo Gonçalves[1,2], António Coelho[1,2]

[1] DEI, FEUP
Rua Dr. Roberto Frias s/n 4200-465, Porto, Portugal

[2] INESC TEC
Rua Dr. Roberto Frias s/n 4200-465, Porto, Portugal

jpulisses@gmail.com, refg@fe.up.pt, acoelho@fe.up.pt

Abstract. One of the biggest pitfalls digital games have is the lack of replay-ability and consequently a limited lifespan. The reason behind this problem is that in most cases, when the player reaches the end of a game, there is nothing more to explore. To address this issue many methods have been researched and implemented. These methods are mainly procedural content generation algorithms (PCG), non-linear stories and the ability for players to create their own content. The problem however remained as the integration of this type of content and their quality is not easy or not done as desirable, especially narratives. Due to the advantages PCG algorithms, they are being integrated into project Orion, a serious game, along with a narrative that aims to create a versatile tool independently of the programming language. This promotes replay-ability as more content can be generated and will help students in their studies since the game may last longer, as well as teachers to use this tool dynamically according to what they want to teach. In addition, since narrative is also an extremely important component of game content, an XML schema has been created to store static stories that will be integrated into the generated game levels seamlessly. As a result, Project Orion will show the implemented dungeon generation algorithm and narrative integration and how it is possible to generate interesting multi-level dungeon games that incorporate all elements in a narrative world which will arch for the player to follow in a 3D virtual world where any content can be easily added, even by people that do not have programming skills.

Keywords
Procedural Content Generation in Games; Serious Games; Dungeon Crawler; Level Generation; Dynamic Generated Content; Narrative

1. Introduction

Serious games are becoming an alternative method for imparting knowledge and skills, both in academic and enterprise environments. This technology, as the name implies, uses game characteristics such as rules, challenges, rewards, interaction and feedback, and wraps them together to solve a problem. By achieving a balance between fun and the problem addressed, these types of games can become a motivational force to engage learners to learn the desired knowledge and skills.

In spite of the intrinsic and extrinsic motivational factors inherent to any game, whether serious or not, one of the biggest engagement pitfalls has always been their longevity and replay-ability. Once the novelty wears off, players rarely feel the need to come back, especially if it is a single player game. This is where the procedural generation of content in Games (PCG-G) may help, by generating content in real-time, either as the game evolves or every time a new game is initiated. Of course content has a very broad meaning, however in this work we are particularly addressing the generation of game levels or maps and integrating it with content that may be generated in multiple ways even external to the game, such as the narrative. The inclusion of narrative in the game is also extremely important since storytelling has always been a crucial element in human development [11], and also functions as an intrinsic motivational factor [10].

In addition to the replay-ability potential that PCG algorithms may introduce, there is also a great advantage of reducing the workload on the design teams by generating to a point where artists can pick the generated assets and polish and add further detail, enriching the overall quality of the application and reducing the time and cost it would take if content was done from the beggining [1]. Doing integration between the generated content and the narrative is a problem, but with methods similar to Natural Language Processing it is possible to achieve this, but also have the type of interaction between the writer directly to the program that allows development teams to save time.

This paper is organized as follows. Section 2 defines the context of PCG and presents the problematic of narrative integration and the proposed solution, integrated with all other generated content showed in the following sections for a serious game for the teaching of computer programming [2]. In section 3 narrative integration is explained and how it is achieved, methods will be shown such as the use of an XML file to describe a narrative so that the generated levels tell a story, as well other content.

In section 4 it is given a context of PCG base algorithms for dungeon generation and the algorithm used is specified thoroughly showing it is strong points and why it was used. In section 5 the dynamically-generated world in Project Orion with all the results from all the integrations, showing advantages of utmost importance in the way it was made allowing not only better results, better team integration and other advantages, this section also includes demonstration. Finally, the conclusions and future work are presented in section 6.

2. State of the Art

This section will begin to explain what procedural content generation in games is, then what type of content can be generated as well as the distinction between adaptive and non-adaptive PCG algorithms. Finally, a few PCG algorithms will be explored.

Procedural Content Generation in games (PCG-G) can be defined as the automatic creation of content, which is achieved through the implementation of an algorithm tailored to a specific end [3]. This process can be achieved through a random process or through deterministic reconstruction based on parameters which will create the

same object every time. Since the main focus of this article is the creation of continuous new experiences in game environments, the focused PCG-G algorithms will have some degree of stochasticity.

Even though content is intended to be greatly different in every generation iteration to avoid the feeling of repeating in the player, it is a good policy not to be completely random. Instead PCG-G algorithms should implement some degree of stochasticity that adheres to constraints, which can be affected by associated parameters, of what content can be generated and how [4]. By generating content randomly within that range of constraints, the result will be adequate and within the scope of the game. Furthermore, by allowing the player to act, directly or indirectly, upon the parametrization of the algorithm the results will be tailored to the choices he made [5]. These algorithms fall under the category of adaptive-PCG and can be useful, for instance, to adjust the difficulty of a generated level to the player's proficiency.

But what type of game content can be generated by PCG algorithms? Theoretically, almost anything can be generated content. Hendrikx et al. [1] defines four classes of game content that can be generated procedurally:

- Game Bits include textures, item properties, sounds, vegetation, buildings, behavior, weather and natural elements.
- Game Space pertains to the environment present in-game, whether they are indoor (dungeons; rooms; houses; etc.) or outdoor (forests; space; underwater; etc.).
- Game Systems such as ecosystems, road networking, urban environment development and entity behavior management.
- Game Scenarios define puzzles; story and level concept.

Another important aspect to take into consideration is whether PCG-G algorithms are run online or offline [3]. A PCG-G online algorithm is usually executed in run-time, as the game progresses, generating new content on the fly. For example Minecraft[1] is continuously generating the environment as the player moves into uncharted territory.

Online algorithms may be adaptive and have two critical requirements. The first of these requirements is speed of execution because if these algorithms take too long doing the necessary computations it will have a detrimental effect on the player's immersion. As for the second requirement, the algorithm needs to ensure that a minimum level of quality is attained.

On the other hand, offline algorithms are usually used during development and subject to specific changes to meet the desired results and as such, these are almost always non-adaptive.

3. Narrative Integration

[1] https://minecraft.net/, 2014

Having a new map to explore every time we play a particular game is without a doubt exciting (at the least the first few times) and can be a source of motivation for the player to explore the game once again. This effect can be significantly amplified if the map also presents a story for the player to follow. But since the maps keep changing every new playthrough, it becomes necessary to have some mechanism that allows the new map to contain the narrative content and plot points that compose the story. The game Diablo[2] is fairly renowned for achieving this effect - while the maps keep changing every new playthrough, the storyline (quests), key objects, places and characters remain the same.

In order to complement the generated maps with a story we had to somehow store the narrative information, as well as the programming exercises associated with each map/level. By storing this information we could then adapt our map generation to match the needs of the story as well as of the exercises.

To this end we created a data format based on XML, as shown in figure 1. This format allows the definition of each of the individual levels, associating with them dialogs, descriptions and objectives. These objectives are associated with exercises, the level they unlock (useful to establish links between levels) and other information related with the narrative and gameplay mechanics, such as object collection, enemy encounters, etc.

The exercises definition is fairly simple but requires that the identity field matches the ones in the exercise evaluator module, for more information about this module refer to previous work [2].

Finally, to give story more life, characters and dialogs are elements also present in the format, although at this stage in a very rudimentary form. Through these tags are the main character, player controlled, as well as other non-player controlled characters can be defined and have dialogs associated with them.

[2] http://eu.blizzard.com/en-gb/games/d2/ , 2014

```
- <game>
    <leveldefinition numrooms="" startlevel="" />
  - <levels>
    - <level idlevel="" keyobjecttype="" description="" iddialog="">
        <objective type="" idlevel destination="" idobject="" idexercise=""
          enemytype="" numenemies="" iddialog_complete="" />
      </level>
    </levels>
  - <objects>
      <object idobject="" name="" />
    </objects>
  - <exercises>
      <exercise id="" active="" difficulty="" description="" iddialog="" />
    </exercises>
  - <characters>
      <character id="" name="" type="" />
    </characters>
  - <dialogs>
    - <dialog id="">
        <entry idcharacter="" text="" />
      </dialog>
    </dialogs>
</game>
```

Figure 1. Base XML format for storing narrative and exercise information.

As this stage of development the dialogues are static and do not allow the player to have different choices (though they can explore in different ways and do any order of exercises), but still means that the story itself is completely linear, however like any content here, it can be changed and improved by the writer having low or none implementation costs and allowing the player to make decisions.

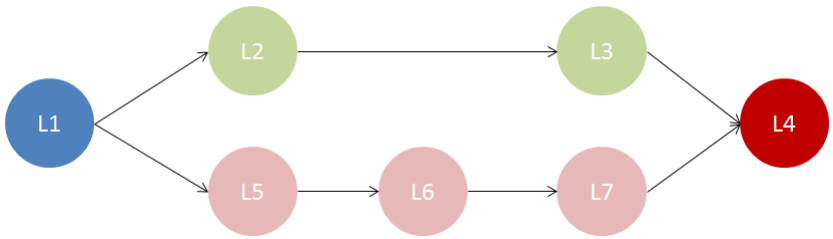


Figure 2. Example of a narrative arch.

4. PCG-G Algorithms for Dungeon Generation

Dungeon level is a instance of a video game level, this instance is often called dungeon because the player has to explore it in order to complete that level.

Game level generation usually have specific constraints that are associated with the type of level being generated which usually require a specific oriented algorithm. In this sense an algorithm for generating an open world with oceans, mountains, etc.

may not be the most adequate to generate a dungeon area. For this reason, in this section, some base algorithms will be presented which are mainly oriented to generate dungeon levels.

4.1. Random Room Placement

The algorithm used consists mainly in populating a designated gridded area with rooms of random size and connecting them through corridors. As parameters this algorithm can receive the number of rooms to generate, the width and the height ratio.

The general workflow of this algorithm is as follows: firstly it starts by parsing the parameters, proceeding to a loop state where rooms are generated. For each room generation loop iteration, the generated room must be first validated before placing it on the map. For example if the room overlaps another or is outside the bounds of the map it is discarded and the room is redone. This process loops until all rooms have been generated or until some other exit condition is met, such as reaching the maximum number of attempts. After all rooms have been successfully generated the next step is the creation of the corridors. This step usually involves using a path finding algorithm such as A* to find the closest path from room to room. When all rooms have been connected, if there are still unconnected doors in some of the rooms, additional dead end corridors may be added or like in Orion non-openable doors can be generated here. Orion also creates doors in spaces between rooms and corridors that are not walls.

4.2. Space Partitioning For Room Placement

In the previous example, during room generation and map placement, it could occur with some frequency that rooms would get discarded. With space partitioning algorithms, such as BSP trees and quadtrees [6] this problem is circumvented because rooms are placed in the empty generated areas and therefore do not overlap other rooms.

BSP tree random placement starts by sub dividing the initial map area into smaller sections by choosing a horizontal and vertical direction. This sub-division is done for *n* iterations. After reaching the n iterations rooms are inserted into the subdivided areas. Note that rooms must be of the same size or smaller than the regions they will be placed on. After all rooms have been successfully placed, the next step is connecting the rooms by looping through all split regions and connecting each immediate neighboring regions ensuring that all rooms get connected.

Similarly to BSP trees, quadtrees start to divide a large area but in this case it divides it in four equal squares. For each new area it is again subdivided until either an area reaches the minimum room size or if the stochastic element of the algorithm stops subdivision. For each leaf a room is placed such that it is entirely contained within the leaf area.

4.3. Cellular Automaton Method

While the two previous algorithms are oriented to generate room-like dungeons, a cellular automaton algorithm is oriented to generate cavern-like areas (closed areas

with a opening) [7]. The first step in this algorithm is to fill an area randomly with walls and empty spaces. Subsequently the algorithm parses each of the grid cells and applies the 4-5 rule: a cell *C* at position *p* becomes a wall if at least five (including itself) of the eight (diagonals included) immediate connected neighboring cells are walls. Parsing each of the cells need to be done simultaneously instead of parsing one by one and using the value of the previous to calculate the next. The reason for this is purely to make it will look less machine processed. One of the problems this algorithm has is that it tends to have very inconsistent results such as wide open areas or disjoint areas. This problem can be solved however by adding an additional refinement rule before applying the initially defined rule[3] . This new rule contains the first one plus: a cell *C* becomes a wall if two or less cells within 2 steps from *C* are walls. A mathematical notation of these rules can be for example:

Rule 1: $W'(p) = C_1(p) \geq 5 \; or \; C_2(p) \leq 2$ Rule 2: $W'(p) = C_1(p) \geq 5$

Where $W'(p)$ represents if there is a wall in a particular position *p*, and $C_n(p)$ represents the number of cells within *n* steps of the position *p* that are walls.

4.4. Algorithm Developed for the serious game Orion

The developed algorithm generates different unique levels for each player each time it is executed. It is a requirement that maps are generated at game run-time.

This algorithm is inspired on a random room placement approach used in the game TinyKeep[4], however with some important differences such as disconnected rooms and multi-level maps. The reason behind this choice was strongly due to the aesthetic look and topology that could be achieved with this algorithm, which feels more varied and less artificial, consequently having more immersion potential which is one of the main goals of the serious game this algorithm [9] is meant to integrate.

As many other algorithms this one also requires some entry parameters to function. These parameters are the number of rooms to be generated, a set of values representing the mean and deviation parameters for a Gaussian distribution function, a maximum room size ratio, the minimum room size and finally the minimum boundary (space) between rooms. These parameters will be further explained as the steps involved in this algorithm are detailed.

The algorithm is divided in different sequential steps that use the data from the previous step to feed the next until the final result is reached. The first step is randomly creating rooms and positioning them over the map area without worrying about overlapped rooms. The room creation is basically a loop that generates as many room as set in the entry parameters. For each room generated, before accepting it, it is verified if it meets certain constraints such as room ratio, to avoid very narrow rooms for example, or if the room has a minimum size, for example to avoid single cell rooms. Rooms random generation is done by using a Gaussian distribution function to ensure that most rooms are within the same size/ratio range and only a few are bigger,

[3] http://pixelenvy.ca/wa/ca_cave.html, 2014

[4] http://tinykeep.com/, 2014

creating a more natural distribution and avoiding having a map filled with only huge rooms.

The next step on this process is to ensure that no rooms are overlapped and to achieve this, a simple separation steering function is used on each room until no room overlaps. This steering process uses the entry parameter minimum border, which can be equal or higher than zero, to increase (or use the default room size if it is zero) to check if two rooms overlap.

The main reason behind this parameter is to prevent rooms to be closely packed, allowing space for later corridor generation. Also, the separation steering process ensures that the room is snapped to the grid to ease the corridor generation. Because this is a multi-level map (each level is a game level which can have two floors), when a new level is added to a generated map in a previous iteration of this algorithm process, some rooms will be selected as level transition rooms.

The selection of these transition rooms is based on their area (the higher the area the more chances it has to be selected as a transition room) and they are meant to establish a transition point between the current level and the next one.

In figure 3, these rooms can be seen as blue if they are transition rooms to the next level, or pink if it is a transition room to the previous level. Upon creating a new map, the previous map transition rooms are passed and integrated within the selected rooms of the new level. These rooms remain fixed during the process of separation steering of the new level to ensure they retain their relative position to the parent's level.

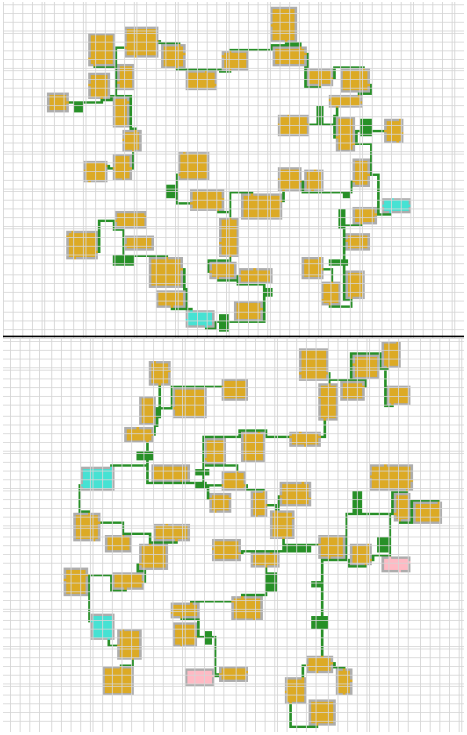


Figure 3. Example of a generated map with two levels. Blue rooms are transition rooms to the next level and pink rooms are transition rooms to the previous level.

With a list of selected rooms the next step is the creation of a graph that connects rooms together. This is achieved by constructing a graph similar to a relative neighborhood graph, except that this one was done in such a way that in some cases a map could have disconnected sub-graphs which could be a potentially way to allow some kind of teleportation mechanic in the game. This can be disabled by simply selecting the two closest nodes of each closest sub-graph and connecting them ensuring this way that every room is reachable by normal means. One of the side effects of this approach was that some rooms were connected by several nodes, creating multiple paths between nodes and to solve this, a minimal spanning tree is generated from the initial graphs, reducing the number of connections between nodes.

The final step is connecting rooms using the graph. To achieve this, a path finding algorithm was used, more specifically A*, which is used to create a route for each graph edge.

To avoid non-upright (zig-zag) corridors the algorithm punishes heavily direction alteration, ensuring that corridors have a straighter format. Also the different cost values are given to cells to ensure that the algorithm will always favor existing corridors, adding more realism and variety to the map, like intersections. Finally, after

a path as been found between two rooms, this path is checked against all non-selected rooms and if an intersection occurs, the room is integrated as part of the corridor and no longer being a room adding further detail to the general map layout (so corridors are not always lines connecting rooms).

5. Project Orion and Results

Project Orion[12][2] as shown in figure 4 is a roguelike[5] game, having the main characteristics such as the fact that Orion generated procedurally as shown in figure 5 and 6, the objects have different names and descriptions, and the game was meant to play alone (though it is easy to implement multiplayer because of the way it was dynamically generated by just adding one more player and create a connection between the players).

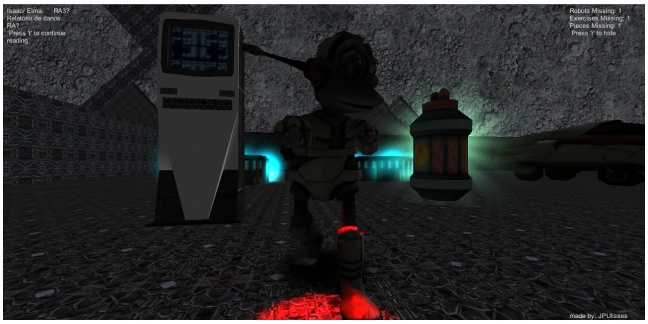


Figure 4. Orion, with different objects to pick and interact.

This project used the algorithm presented in section 4 to generate the map, that information was used to draw 3D elements such as rooms, corridors and placing doors.

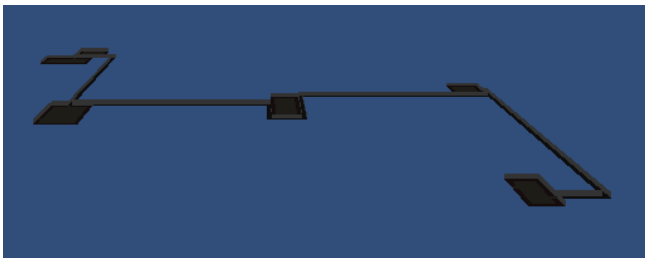


Figure 5. A generated level with a few rooms and doors.

[5] http://www.roguebasin.com/index.php?title=Berlin_Interpretation, 2008

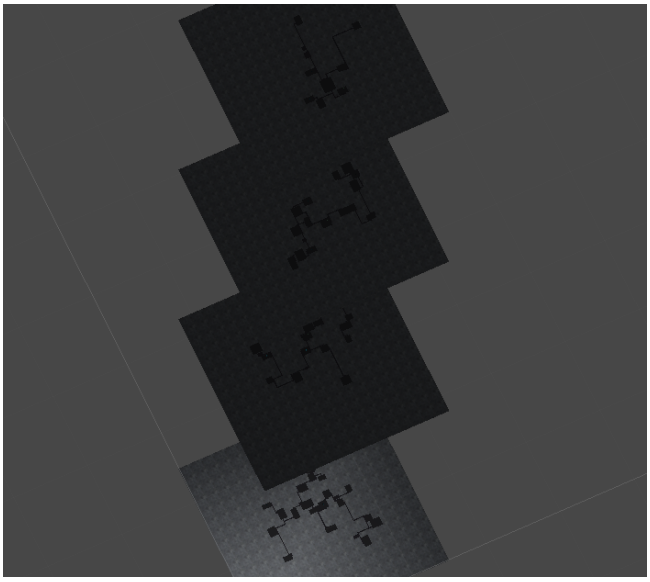


Figure 6. The generated levels in Y axis, each with more rooms than the example given before. The player teleports from one level to another after completing the objectives, physical stairs could be added if they had fit into the theme of the game.

In the game most content is present and editable in a external XML file, such as gameplay mechanics, objects, enemies, programming exercises and the narrative itself. These allow to dynamically create a 3D world with content to play and interact with it in different ways, allowing the player to explore and solve the problems presented. The player can make use of the dialog following the events and the conversation between characters or be guided in the right way, the programming exercises can also offer tips and help. This content can be seen, and edited externally in the demonstration[6].

The way it was programmed dynamically and using Unity3D, allows any content to be replaced at any time by referencing another content. This also allows the content to be adapted to different programming languages. Thanks to Unity3D, these capabilities allow teams to work independently and put together their work at a later stage, and also constantly update and improve it, with low cost of implementation. For example in the demonstration, some objects such as the enemy robot that comes from Unity3D Platform Tutorial, have a JavaScript code, however it is instantiated and asked to be destroyed by C# code.

[6] https://feupload.fe.up.pt/get/grRy5yI2kcOcK0a, 2014

6. Conclusions and Future work

From the aesthetic and functional point of view, the generated maps achieve their goals, to generate dynamic content as shown, however it is possible to go further and not only use this to speed the process of generating large content, but also allow multi-disciplinary teams to work together with ease and update their work on later stages of development with low implementation cost. The extra capability of the platform used (Unity3D) to instantiate and communicate with objects of other programming languages enhances this capability even further.

PCG lay the foundation for a rich dungeon environment and ensure that there are several pathways that can be taken to reach the end of each map (level). Coupled with the XML narrative description file, these maps are brought to life which acts as motivation factor for the player to explore the map and discover the story, consequently having to solve the proposed exercises.

Everyday the games that use PCG methods require people from multidisciplinary areas, that not only must know how to program, but they must know the generated content in order to predict it, limit the way it will be generated if needed and generate it in a pleasant way for the user or with some degree of expected quality. This claim should be changed as it is not always true as shown, teams can work separately, each with different areas of knowledge, however the team in charge of putting all together, generating the content based on the work of the other teams, must have some knowledge of that area in order to create the best result.

For the narrative at this stage it is still done manually, however in the future we hope to generate the stories and provide a complete 'new' experience each playthrough, as well as varying exercise parameters so that these are also unique to each student. Additionally the XML format needs to be further enhanced to allow player dialog choices, multi-dependent objectives, non-player character interaction, non-serious side quests and elements, events, locations, etc. However at this stage it is possible to have player-created content as shown in the demonstration, while this exists for entertainment games, it is harder to do on serious games as we did, this can create new experiences and promote experience sharing between the students.

The algorithm for dungeons is not finished however, for the moment all rooms are rectangular which may suit most cases, but to increase immersion some rooms could be improved, for example, by using predefined models that add more detail to the map. Another evolution is to have a predetermined pathway that is passed to the algorithm and generate around it a map level. This could be useful in the eventuality of wanting to define a set of rooms with specific exercises and narrative plot points.

Preliminary results show that most players pointed out it was easy to get to play and to play Orion. This information includes the game itself, how user friendly it was, but the whole process since downloading the game, which thanks to Unity3D makes the process of building the project much easier, creating a executable file ready to be played. This also shows the game is ready to be tested, however more content is needed for further experiences, these can done either by professors or students if they wish to do so.

Acknowledgements

"The Media Arts and Technologies project" (MAT), NORTE-07-0124-FEDER-000061, is financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT)."

References

1. Hendrikx, Mark, et al. "Procedural content generation for games: a survey." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 2013.
2. Coelho, António; Kato, Enrique; Xavier, João; Gonçalves, Ricardo: Serious Game for Introductory Programming. In Proceedings of the Second International Conference on Serious Games Development and Applications (SGDA 2011), Lisbon, 2011.
3. Togelius, Julian, et al. "Search-based procedural content generation: A taxonomy and survey." Computational Intelligence and AI in Games, 2011.
4. Togelius, Julian, et al. "What is procedural content generation?: Mario on the borderline." Proceedings of the 2nd International Workshop on Procedural Content Generation in Games. ACM, 2011.
5. Georgios, Yannakakis and Togelius, Julian. "Experience-driven procedural content generation." Affective Computing, IEEE Transactions, 2011.
6. Togelius, Julian; Shaker, Noor; Nelson, Mark J. "Procedural Content Generation in Games: A Textbook and an Overview of Current Research", 2014
7. Johnson, Lawrence; Georgios, Yannakakis; Togelius, Julian. "Cellular automata for real-time generation of infinite cave levels." Proceedings of the 2010 Workshop on Procedural Content Generation in Games. ACM, 2010.
8. Valtchanov, Brown, Joseph. "Evolving dungeon crawler levels with relative placement." Proceedings of the Fifth International C* Conference on Computer Science and Software Engineering. ACM, 2012.
9. Doull, Andrew. "The death of the level designer." Internet: http://pcg. wikidot. com/the-death-ofthe-level-designer, last accessed in 2008.
10. Pintrich, Paul R., et al. "Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ)." Educational and psychological measurement, 1993.
11. Schank, Roger C., and Robert P. Abelson. "Knowledge and memory: The real story.", 1995.
12. Ulisses, João; Coelho, António: "Solução de geração procedimental de níveis de jogo", Oporto, June 2014.
13. Togelius, Julian, et al. "Procedural Content Generation: Goals, Challenges and Actionable Steps", 2013.

SESSION 4

VIRTUAL SIMULATION

A Multi-player Approach in Serious Games: Testing Pedestrian Fire Evacuation Scenarios
Marcos André Oliveira, Nelson Miguel Pereira, Joao Emílio Almeida, Rosaldo Rossetti and Eugénio Costa Oliveira

3D Simulation Environment: Education and Training
Hugo Barbosa

A Multi-player Approach in Serious Games: Testing Pedestrian Fire Evacuation Scenarios

Marcos Oliveira[1], Nelson Pereira[1], João E. Almeida[2,3], Rosaldo J. F. Rossetti[2,3], Eugénio Oliveira[2,3]

[1]Mestrado Integrado em Engenharia Informática e Computação (MIEIC)
[2] Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
[3]Departamento de Engenharia Informática (DEI)
Faculdade de Engenharia da Universidade do Porto (FEUP)
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
{ei09149, ei09025, joao.emilio.almeida, rossetti, eco}@fe.up.pt

Abstract. Serious Games are being increasingly used as a tool for various applications, including social simulation. One of the domains of application is fire safety training and behaviour elicitation. Injuries and fatal casualties having its origin from fires are a concern for fire safety engineers, building managers, and emergency responders. Training building occupants has been a challenge for quite some time. One approach is by classroom education or a more practical way by performing exercises, called fire drills. The former consists of a method for validating evacuation plans and tests what-if scenarios. Another important issue is the lack of human behaviour data; this aspect is often referred to as a drawback to evacuation simulator developers. The elicitation of behavioural knowledge to feed simulators constitutes a critical aspect for which some researchers have proposed the use of Serious Games. This paper addresses these issues in respect to: i) train escape procedures using Serious Games as an educational tool; ii) acquire valuable knowledge using the concept of participatory simulation. A test-bed using Serious Games, developed on the Unity3D framework, implements a multi-player approach taking advantage of its features. The experimental setup is presented and some test results are discussed. Future work is two-fold: expand and refine the scenarios for a wider set of possibilities; perform massive data collection that will be used to feed existing multi-agent evacuation simulators.

Keywords: serious games, fire safety education, evacuation, multi-player, behaviour analysis.

1 Introduction

Fire is considered to be the most dangerous hazard for building occupants [1]. Every year many casualties due to fire occur, some of them resulting in death. A recent example was the fire at the Brazilian discotheque "Kiss", January 27th 2013, resulting in death of 242 people, many of which college students. Such occurrences are many times due to the lack of information and training of the occupants as well as the

emergency responders. So, training and education are the best solution to minimize the number of casualties that each year claim the lives of many around the world.

Although fire drills and classroom learning techniques are a possible way to teach fire safety, the use of Serious Game (SG) has been proposed as a good method for training and education [2]–[5]. Indeed, video games has fostered the implementation of SGs with diverse goals other than entertainment [6].

The use of Virtual Reality (VR) based applications for both simulate situations that are too dangerous for exposing real people to and as an aid for training and education is not new [7], [8]. SGs are a good trade-off compromise for the development and rapid prototyping of low-cost VR applications [9]. Unity3D is a successful platform used worldwide for the development of video games presenting high quality graphics, animation and VR features [10].

Another issue is the study of the self-organizing processes associated with building occupants when facing an emergency and having to abandon it, of great importance to assess the safety of the building, pre-defining possible scenarios, and to implement emergency evacuation plans. To help researchers and designers, computer evacuation models were developed for testing what-if scenarios. Most of these evacuation simulators are agent-based. For the intelligent agent behaviour modelling, there is a urgent need of real data to validate and calibrate such models [11].

Our team has been developing SGs aiming to train occupants for the evacuation procedures out of a building facing the presence of fire or other hazardous situations, for some time and having some experiments with good results. For this purpose, we have devised a framework coined Simulation of Pedestrians and Elicitation of their Emergent Dynamics (SPEED) for the elicitation of human behaviour in emergency situations. This framework consists of a methodological approach aiming at the elicitation of human behaviour in hazardous situations, and the use of the collected data to breed and grow an artificial society [12].

One important aspect of the SPEED development consists in having participatory simulation, in which some players interact in the same virtual environment. Taking advantage of the multi-player feature of Unity3D, we envisage a test-bed in which players share the same scenario, having to leave as quickly as possible, as soon as the fire alarm sounds. The experimental setup described in this paper aims to train and educate the players in evacuation techniques as well as elicit their behaviour when facing the urgent need of evacuation from a building.

For the sake of demonstration, we setup a scenario consisting of an auditorium, which has been used in previous experiments, this time using various players simultaneously sharing the same web-based application developed under the Unity3D framework. The final goal is to understand how subjects behave and try to extrapolate their emergent behaviour for fire safety planners and engineers as well as pedestrian evacuation modellers.

The remainder of this paper is organised as follows. Section 2 presents some background and related work in the field of fire safety evacuation techniques, SG and VR, behaviour elicitation and participatory simulation. Section 3 is used to introduce the implementation, whereas Section 4 the experimental setup. Section 5 discusses results from the experiments and tasks to accomplish such a realization. Finally some conclusions are drawn and future works, as well as developments are presented in Section 6.

2 Background and Related Work

Before moving further on the explanation of the experimental setup and the results obtained, it is important to introduce some valuable concepts and subjects underlying this project.

Fire safety training and education

The domain of fire safety is utterly dependent on the human behaviour, both in its origin as well as during its development [13]. Particularly the evacuation process in which occupants must leave whatever activities they are engaged to move as quickly as possible towards the safest and nearest exit. It is commonly noted that the human behaviour is of paramount importance for the outcome of such process [14]–[17].

To train and educate building occupants it is usual to perform evacuation exercises also called fire drills.

Fire drills and building evacuation

Fire drills are mandatory in many countries. They are performed to test emergency plans but also to educate building occupants in fire escape procedures. Teaching fire safety skills is an important issue to diminish the number of casualties and increase the level of safety.

Serious Games are a powerful tool to catch the attention of the participants, who otherwise would consider those exercises fastidious and boring, only carried out if forced by some disciplinary obligation (professional, academic or to avoid some kind of penalty). Also, the fire drill or emergency scenario can be more realistic without creating situations of danger to the trainees or building occupants. Data stored by the computational tool can also be used for statistic purposes, or to validate and calibrate computer models.

The Serious Games Concept

Serious Games has gained a great prominence in the field of Digital Games within the last years, by using high-definition graphics and state-of-the-art appealing animation software [18]. It presents a great potential as a tool to be used for other purposes rather than mere entertainment. Applications have a wide range of domains, naturally including social simulation, where data collection of player attitudes can be later used for statistical analysis, and behavioural pattern recognition.

Contrary to the primary purpose of entertainment in traditional digital games, SGs are designed with a more serious purpose with respect to the outcomes reflected in changes to the player behaviour [19].

A game is an artificially constructed, competitive activity with a specific goal, a set of rules and constraints that is located in a specific context [20]. SGs refer to video games whose application is focused on supporting activities such as education,

training, health, advertising, or social change. Freitas [21] has identified a set of benefits from combining SGs with other training activities: i) the learners' motivation is elevated; ii) completion rates are higher; iii) possibility of accepting new learners; iv) possibility of creating collaborative activities; v) learn through doing and acquiring experience.

Other aspects that draw video game players' attention are fantasy elements, challenging situations and the ability to keep them curious about the outcomes of their possible actions [22].

Using Unity3D and Photon Unity Network to implement the Serious Games

The SGs used as example in this research was created using the Unity3D game engine. Unity3D was selected due to its main characteristics: i) powerful graphical interface that allows visual object placement and property changing in runtime (especially useful to rapidly create new scenarios from existing models and assets and quick tweaking of script variables); ii) the ability to develop code in JavaScript, C# or Boo; iii) simple project deployment for multiple platforms including the Web, which makes it possible to run the game on a Web browser, a feature that is particularly interesting for massive data collection. This last aspect is something that we aim to explore in a near future.

The Photon Unity Network framework was used to implement the network component of the game, making the multiplayer implementation much easier as the server is already setup.

Multi-player

Multi-player is a game mode concept in which it is possible for two or more players to play in the same game at the same time. Classically it is used for cooperatively play (team-based games), or head-to-head competition (famously known as deathmatch). There are usually two modes: split screen in which the users play on the same system and share the screen, or via a Network, which can be local (LAN) or on the web via game servers.

3 Implementation

The implementation of the Serious Game, coined EVA[23], was made using the web-based deployment version of Unity3D. For the multi-player game mode, we used the Photon Unity Network (PUN) which creates a set of "rooms" where the players are connected (see Fig.1) via a network, either local (LAN) or over the Internet, through game servers. Fig.2 shows a screenshot of one experiment with more than one player.

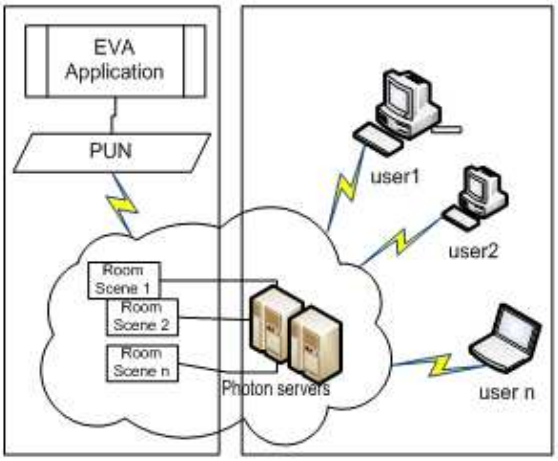


Fig. 1: Implementation architecture

4 Experimental Setup

The Office Room Scene

Fig. 2: Screenshot of Experiment 1 with two players in action simultaneously

This scene consists in subjecting the users to five different case scenarios. Each scenario took place in a virtual office room with the real time interaction of various players. The experiment starts when the player is spawned in the office room and starts hearing a fire alarm. Each player has then to find the exit of the building; however, each scenario had small differences that made each case special.

First scenario

In this scenario, after getting to the door of the office, the user cannot see any emergency exit sign that indicates which direction to take (left or right). The results obtained can shed a light on how people react in these types of situations which lacks any emergency signs is pointing out the right direction (see Fig.3 left).

Fig. 3: Screenshot of a case scenario with no emergency exit signal (left); and with the emergency sign pointing to the left (right).

Second scenario

This time around there is an emergency sign that indicates the user to take the left to get to the emergency exit. This scenario has the objective of studying the level of attention that people have in panic situations (see Fig.3 right).

Third scenario

The third scenario was designed to study the risk factor of the users, meaning that we wanted to analyse what would happen when the users see that the emergency sign communicates that to get to the exit the left path must be taken, but said path is blocked by a cloud of smoke (see Fig.4 left).

Fig. 4: Scenario with a cloud of smoke blocking the exit (left) and a fire (right).

Fourth scenario

This scenario is very similar to the previous one. The objective here is to study what is the user reaction when faced with a dangerous situation. In this case, when confronted with a wall of fire blocking the path to the exit indicated by the emergency sign. This scenario illustrates a case that is simpler to test virtually rather than in the real world (see Fig.4 right).

Fifth scenario

In the final scenario of the first experiment, crowd influence was tested. The user sees that despite the emergency sign depicting the exit to the left, there is a crowd of people running to the right (see Fig.5).

Fig. 5: Scenario with a crowd running in the opposite direction of the emergency exit sign

The cinema auditorium scene

In this scenario, as with the office room scene, we want to study how the interaction between players influences the choice for an emergency exit. However, as this scenario presents itself with a lot more space, we can run the experiments with a larger number of players at the same time. The experiment starts when a player is connected to a multiplayer session, then the player is placed in front of a random chair in the cinema auditorium. A fire alarm is then heard and the player has to find a way out. There are two exit points, the emergency exit, located near the movie screen and the auditorium's entrance where moviegoers enter the room (as we can see in Fig. 6 and 7).

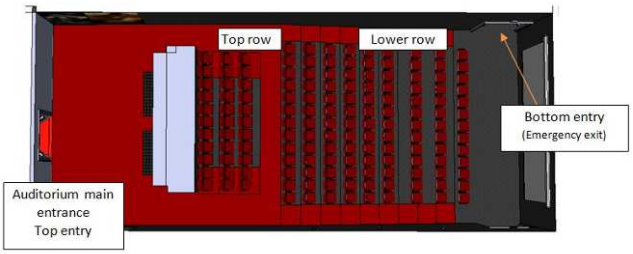


Fig. 6: Screenshot of a schematic representation of the cinema's auditorium

As with any other common cinema auditorium, it is characterised by a certain inclination and steps of stairs, which may cause difficulties in the evacuation process.

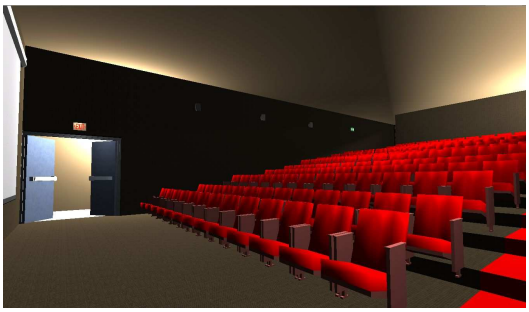


Fig. 7: Another perspective of the auditorium.

The game genre: First Person Player

First Person Players (FPP) are characterised by placing players in a 3D virtual world which is seen through the eyes of an avatar. When playing, user has the feeling of being actually on the location site, moving around, and giving the best possible sensation of immersion.

The controls for the SG presented to the children follow the common standards for the FPS genre, using a combination of keyboard and mouse to move the character around the environment.

5 Data collected and results analysis

In a previous experiment, a group of 19 children from a local elementary school were selected to play the SG. Their main characteristics are presented in **Table 1**. All of

them were used to interact with computers, tablets and video games. In fact, almost all said having at least one game console at home (e.g. Playstation, PSP, Nintendo, Wii).

Table 1. Population sample's characteristics.

Data	Values
Number of subjects	19 (100%)
Male subjects	9 (47%)
Female subjects	10 (53%)
Mean age	7,58
Age SD	0,96
Left-handed	4 (21%)
1st Grade	2 (11%)
2nd Grade	8 (42%)
3rd Grade	3 (16%)
4th Grade	5 (26%)

Preliminary results from the experiments

Most subjects prefer the keyboard+mouse combination (15 – 79%) instead of the joystick (3 – 16%). It was noted that the youngest children (6 or 7 years old) attending the 1st or 2nd grade, had more difficulties to understand the SG concept and to interact with it. The children of the 3rd and 4th grades were more comfortable with the computer commands and the SG aim.

The exit-choice scenario was probably the more challenging of the two role plays and the one that kids enjoyed the most. Only two failed to see the emergency sign pointing left; they confessed that were not aware of its meaning. These were the youngest (7 years old) so it is understandable their lack of knowledge.

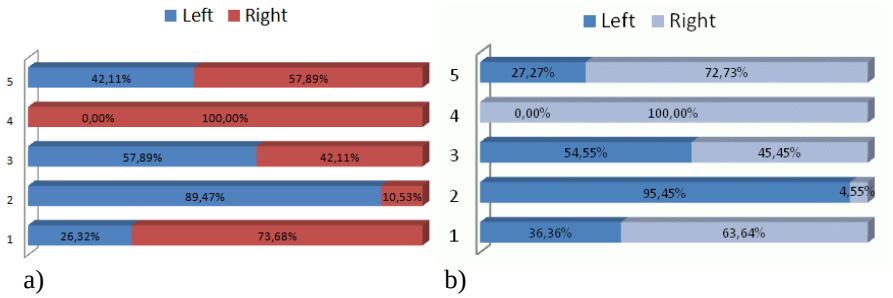


Fig. 8: a) results in percentage for each exit-choice test scene; b) results of a similar test with a population of adults

Fig.8a) shows the results in percentage of the exit-choice test scenario for the children. Fig.8b) shows results from a similar test with a population sample of adults. Comparing both graphs, it is clear that the results are very similar. Only in scene 5, testing the tendency of following others, we realize that kids are more prone to follow

the exit sign (42.11%) than adults (27.27%). Perhaps this fact is due to children are better educated to obey rules than adults. Analytical results are shown in Table 2.

Table 2. Results from the exit-choice scenario.

Scene	Left	Right
1. Tendency to turn left / right	5	14
2.Tendency to follow emergency signs	17	2
3.Tendency to go through smoke	11	8
4.Tendency to go through fire	0	19
5.Tendency to follow others	8	11

Expected results

One of the purposes of this multi-player approach is to analyse and to compare the results with the data collected with the single player version. Added to this, it is expected to extrapolate some players' behaviours when they know they are not alone in the simulation.

So far we have prepared two role play multi-player scenarios for which some experiments can be made in the future, for the purpose of social simulation research.

6 Conclusions and future work

This paper presented the use of SGs to acquire human behaviour when facing the urgent need of evacuating from an unknown building and having to deal with a set of unexpected situations and obstacles. A novel aspect consisted in the implementation of a multi-player component allowing the study of the interaction between players. Results are promising and extremely valuable for fire safety practitioners as well as for evacuation modellers, since data on human behaviour is scarce and much sought-after, particularly for specific groups such as children, elderly or people with disabilities. The knowledge elicited using this methodology might be used by evacuation simulators when trying to model situations.

This method has proved to be useful for data collection as well as training. The subjects that participated in previous experiments, at the end, were taught how they should behave in a real situation, and we are confident that the lessons learned will be hardly forgotten. After the tests, they showed a new confidence and knowledge in fire safety. The multi-player feature, although implemented, was not fully tested at the time of the writing of this paper, due to time constraints and the lack of sufficient concurrent players to test the scenarios. This aspect will be pursued in the future.

Collecting information from the multi-player version of EVA will be the next step of this research. For that, a certain number of users must be chosen as well as observers to collect the data. The massive data collection process using the multi-

player feature developed and presented in this paper will allow us to compare them with results obtained from the single player version of the application.

During the course of the implementation of the multi-player framework, new and interesting features and scenarios arose, that are proposed as future work:

- Having different non-playable characters (NPCs) in the scene, each representing different roles. For instance, having people with disabilities, children and the elderly should cause different reactions on the players. For example, what would the player do if an elderly is passed out on the floor?
- Communication between players would elevate the player interaction even more, bringing the simulation even closer to the real world. However, restrictions would have to be placed so that players could only communicate with others only in a certain range around them.
- The implementation of a visual identifier, such as the players' name above the player's head, for example, would allow the testing of situations in which people tend to follow others they know.
- Registering the amount of time the player has been subjected to smoke (reducing visibility) and even fire, would make the simulation closer to the real world. Dying states and animations would be implemented to show that the user failed to escape. This situation would hopefully cause more stress to the individuals being tested, as it happens in real life.
- Currently the player only moves by walking. By implementing a running feature, this could become a limited resource in the simulation, meaning that the player would become tired over time.
- Implementation of a virtual player, whose objective would be to record the game session, making the data analysis much more efficient.

This work has given us the opportunity to use a complex framework for game development, expand an existing application under development at LIACC, the EVA evacuation simulator, by providing the multi-player capability which will permit new scenarios and data collection opportunities.

The ultimate goal of this research is to provide fire safety engineers and building managers with a tool for training and educating fire safety skills, as well as to grant researchers a means for human behaviour elicitation.

References

1. M. Kobes, I. Helsloot, B. de Vries, and J. G. Post, "Building safety and human behaviour in fire: A literature review," *Fire Saf. J.*, vol. 45, no. 1, pp. 1–11, Jan. 2010.
2. J. F. Silva, J. E. Almeida, R. J. F. Rossetti, and A. L. Coelho, "A Serious Games for EVAcuation Training," in *IEEE 2nd International Conference on Serious Games and Applications for Health (SeGAH 2013)*, 2013.
3. J. Ribeiro, J. E. Almeida, R. J. F. Rossetti, A. Coelho, and A. L. Coelho, "Using Serious Games to Train Evacuation Behaviour," in *CISTI 2012 - 7^a Conferencia Ibérica de Sistemas y Tecnologías de Información*, 2012, pp. 771–776.
4. J. Ribeiro, J. E. Almeida, R. J. F. Rossetti, A. Coelho, and A. L. Coelho, "Towards a serious games evacuation simulator," in *26th European Conference on Modelling and Simulation ECMS 2012*, 2012, pp. 697–702.

5. E. Cordeiro, A. L. Coelho, R. J. F. Rossetti, and J. E. Almeida, "Human Behavior Under Fire Situations – Portuguese Population," in *2011 Fire and Evacuation Modeling Technical Conference*, 2011.
6. J. F. Silva, J. E. Almeida, R. J. F. Rossetti, and A. L. Coelho, "Gamifying Evacuation Drills," in *Third Iberian Workshop on Serious Games and Meaningful Play (SGaMePlay 2013)*, 2013.
7. V. Balasubramanian, D. Massaguer, S. Mehrotra, and N. Venkatasubramanian, "DrillSim : A Simulation Framework for Emergency Response Drills," in *IEEE International Conference on Intelligence and Security Informatics, ISI 2006*, 2006, pp. 237–248.
8. L. Gamberini, P. Cottone, a Spagnolli, D. Varotto, and G. Mantovani, "Responding to a fire emergency in a virtual environment: different patterns of action for different situations.," *Ergonomics*, vol. 46, no. 8, pp. 842–58, Jun. 2003.
9. A. Navarro, J. V. Pradilla, and O. Rios, "Open Source 3D Game Engines for Serious Games Modeling," in *Modeling and Simulation in Engineering*, 2012, pp. 143–158.
10. J. P. M. Ribeiro, "Serious Games Applied to Pedestrian Modelling and Simulation," Master Dissertation, Engineering Faculty of Porto University, Porto, 2012.
11. J. E. Almeida, Z. Kokkinogenis, and R. J. F. Rossetti, "NetLogo Implementation of an Evacuation Scenario," in *Fourth Workshop on Intelligent Systems and Applications (WISA'2012)*, 2012.
12. R. Rossetti, J. E. Almeida, Z. Kokkinogenis, and J. Gonçalves, "Playing Transportation Seriously: Applications of Serious Games to Artificial Transportation Systems," *IEEE Intell. Syst.*, vol. 28, no. 4, pp. 107–112, 2013.
13. S. Horiuchi, "An Overview of Research on 'People-Fire Interactions,'" *Fire Saf. Sci.*, vol. 2, pp. 501–510, 1989.
14. B. B. Pigott, "Fire Detection and Human Behaviour," *Fire Saf. Sci.*, vol. 2, pp. 573–581, 1989.
15. E. D. Kuligowski, "Guest Editorial: The Significance of Pedestrian and Evacuation Dynamics," *Fire Technol.*, vol. 48, no. 1, pp. 1–2, Jul. 2011.
16. E. R. Galea, "Evacuation and Pedestrian Dynamics Guest Editorial – 21st Century Grand Challenges in Evacuation and Pedestrian Dynamics," *Saf. Sci.*, vol. 50, pp. 1653–1654, 2012.
17. M. Kobes, I. Helsloot, B. De Vries, N. Oberijé, and N. Rosmuller, "Fire response performance in a hotel. Behavioural research.," in *Interflam 2007 - 11th international fire science and engineering conference*, 2007, vol. 2, pp. 1429–1434.
18. J. E. Almeida, J. Tiago, P. Neto, B. M. Faria, R. J. F. Rossetti, and A. L. Coelho, "Serious Games for the Elicitation of Way-finding Behaviours in Emergency Situations," in *CISTI 2014 - 9ª Conf. Ibérica de Sist. y Tec. de Información*, 2014.
19. A. Frey, J. Hartig, A. Zinkernagel, and H. Moosbrugger, "The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting," *Comput. Human Behav.*, vol. 23, no. 4, pp. 2026–2039, 2007.
20. R. Hays, "The effectiveness of instructional games: A literature review and discussion," Orlando, Florida, USA, TECHNICAL REPORT 2005-004, 2005.
21. S. I. de Freitas, "Using games and simulations for supporting learning," *Learn. Media Technol.*, vol. 31, no. 4, pp. 343–358, Dec. 2006.
22. J. Kirriemuir and A. McFarlane, *Literature Review in Games and Learning*. Future Lab Series, Report 8, 2004.
23. J. F. M. Silva, J. E. Almeida, A. Pereira, R. J. F. Rossetti, and A. L. Coelho, "Preliminary Experiments with EVA - Serious Games Virtual Fire Drill Simulator," in *27th EUROPEAN Conference on Modelling and Simulation (ECMS 2013)*, 2013.

3D Simulation Environment: Education and Training

Hugo Barbosa

Doctoral Program in Informatics Engineering,
Faculty of Engineering, University of Porto, Portugal
pro11027@fe.up.pt

Abstract: The advances in the interaction and visual simulation environments related to the declining cost of computers and the constant increase in the processing power, have enabled significant progress in how to interact in these environments, allowing their greater use in the analysis of real situations and as a tool for acquiring knowledge and supporting decision making. This article aims to analyze 3D simulation environments for educational purposes and to highlight the usefulness for people with special needs. The technology, combined with appropriate interactivity and visual environments, including 3D simulation can be an asset in the teaching-learning process. This paper addresses the issue by exposing case studies and features results on the behavior of students in the experience with 3D simulation environments.

Keywords: Interactivity, Simulation, 3D environment, Education, Training

1 Introduction

The technology follows the life of people showing a clear presence in their daily lives and it is also getting increasingly affordable. The adjustments as well as the technological progress are constant making new ideas, which result of the creativity and the imagination that follow the modern times, be possible.

The use of simulated environment provides interactivity and experimentation, an enabler of knowledge and analysis of various situations. Thus, among the various areas, the use of these environments in the learning process can provide greater motivation, helping the resources used in the traditional method of teaching. The models which tend to use new technologies seek simple visual environments and easily understandable by the target audience.

The 3D visual simulation is an ever-growing area, proving to be an instrument of support in various areas. Therefore, the aim of this report is to analyze, propose a prototype and explore the impact of 3D simulation environments at school in particular for people with special needs.

The main goal is to provide an idea of the tasks required to perform, allowing the interaction and the visualization of the different methods to achieve them. Although it is simulated, the environment presents a real situation, with the advantage of containing tools to help understanding the task. A greater challenge is to make the content appealing and visually stimulating.

The use of 3D graphical user interface has been proved useful in the learning methods, such as the computer-aided design (CAD), facilitating the development of the project, understanding and further education.

The analysis of the existing resources allowed the knowledge of the current spectrum of existing models which in conjunction with the needs survey conducted in this paper have helped to understand the shortages. Thus, based on this acquired knowledge, a proposal for a simulation model was made. The current available simulators for teaching present a user interface, which is sometimes less obvious as it is the reflection of several options, ultimately confusing and diverting the attention from the real objective and making handling complex. Therefore, the students usually find it difficult to work with, which makes the learning process a long and demanding task. To aid in this task, there is a study based on prototypes being developed in order to provide greater motivation and interest by the public. Features such as animations and three-dimensional statistical graphs are being thought of as options to facilitate the perception of the operation.

2 State of the Art

The current technology allows find a wide variety of simulation environments for different scenarios in education, for example: simulations in business education, simulation in assembly of computers and similar situations, training simulation in cars, plane and many others, but many of them do not take into account specific needs of particular users. Sometimes these environments are important tools with a strong impact on the development of these users that show difficulties at various levels and it can limit others practices in real context. Transforming students from passive observes of linear material to active operators of interactive content, these 3D simulation environments allow students to become immersed and involved.

One example of a tool the simulation environment that simulates the assembly of computers is the Cisco IT Essentials Virtual Desktop. It has a set of options which allows exploring the components, verifying knowledge or watching a demonstration. A different example is the design educational Game Based Simulation of ForgeFX to teaching players about everyday solutions.

Training and education have really changed in recent times. They have moved from classroom training to a continuously model where people learn at any time [1].

Increasingly new 3D technologies are being called on to enhance simulations, using various methods such as modelling, digitization and virtual reality to create high

quality and realistic digital content. This allows learners to experience resources that aren't normally available to them, whether it be due to their being in an inaccessible location or the element being in a fragile condition [2].

3 Simulation Environment

The diversity of applications involving simulated environments is growing. The simulation environments appear as a way to assist and facilitate the knowledge, to experience and analyze the different models. The increase of the use of this method entails some precautions derived from the complexity that they present due to the interaction. The understanding and the interaction with the system by an inexperienced user may not be a simple task since the selection, the manipulation and the orientation of objects may not be intuitive and so it is difficult to adapt.

Training and educational simulations are created in order to facilitate students or users learning. Because of its purpose, an educational simulation is an abstracted representation of the target system, which tries to show the complexity and realism of the element. In [3] have divided educational simulations in two main categories: operational simulations and conceptual simulations. Operational simulations are designed to facilitate the construction of practical knowledge, for example, in areas such as training. Conceptual simulations, on the other hand, are designed to facilitate conceptual knowledge construction on the part of the students. They are based on conceptual models, used within subject domain education, which simulate the relationships that exist between the variables of a real world system, while at the same time allowing the user to manipulate those variables [4]. The applications analyzed in this study are in both categories.

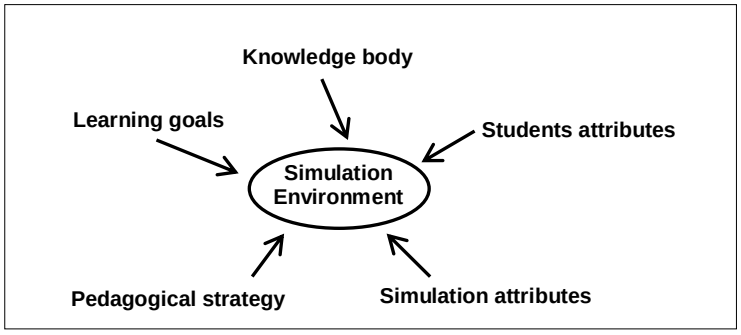


Figure 1 - Aspects of simulation environment training and education

In [5] "Interactive Storytelling: techniques for 21st Century Fiction", A. Glassner discusses what he calls the myth of interactivity, explaining that in recent

years the concept of interactivity was placed in a very high level, because even with the biggest interactivity possible if the environment where the user is interacting is not interesting, the immersion will not succeed. In this sense, one can find various applications that even holding good interactivity, were unsuccessful because they did not captivate the user's interest in exploring it.

Similar discussion can be taken into account in the context of education, where various softwares created for this purpose only reproduce the concept of the paradigm to be transmitted. In this case, we have to think not only about the user's interaction with the environment, but in the interest and the curiosity he may feel in exploring the environment offered.

There are several simulation tools available and within reach of a simple search on Internet. In general they have a set of options, which allow the user to know the components of the area, to check knowledge or to watch a demonstration.

The technique for selecting objects emerges as an important element in the performance of the user when interacting with the system.

4 Implementation and evaluation of results

When you want to build a three-dimensional environment with certain characteristics to be inserted into a product of education, a set of specific tools in the area of 3D is needed. Currently, we can find on the market multiple offers of 3D design software for free and also commercial versions that allow three-dimensional structures to develop from small to complex 3D environments [6].

The development of the whole simulation environment (Figure 2 and 3) presented here, was concerned with the need to create a simple and intuitive interface so that you can focus your attention on the systems studied and not on the learning of the use. In this case, the simulation environment to computer assembly, appear as a way to assist and facilitate the knowledge, experience and analyze the different models. The highlight of this simulation is derived from the complexity that it presents, due to the interaction.

Figure 2 - Interface of the simulation

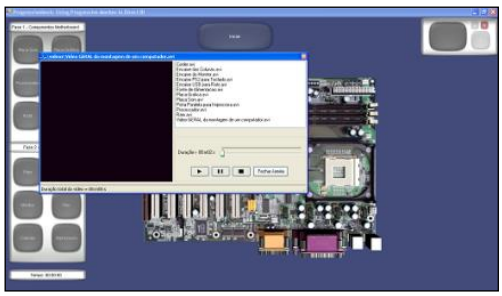


Figure 3 - Video help

The usual method in the evaluation of the 3D interaction is the users test, task-based. Therefore, in this experiment a group of users (32 students) from a local high school were selected to test 3D Simulation Environment in computers. Their main characteristics are presented in Table 1. This group interacted with 3D simulation environments making it possible to see their behavior regarding this application.

Data	Values
Number of subjects	32
Male subjects	25
Female subjects	7
Minimum age	14
Maximum age	23
Mean age	16,4

Table 1 - Students samples characteristics.

This is a stimulating environment for users concerning the visual level, based in the trend perceived in the context of training in the classroom. However, the interactive element appears to be a difficult component, because the features are leading many users with fewer skills to need assistance in completing the task. The possibility of demonstration is an enriching factor, according to the analysis done in the same context.

The applications analyzed in this study revealed similar points of view. Thus, the point of view of the observer is fixed, a strategy which saves the user from

handling complementary options at the same time that loaded the environment with options. However, in accordance with the results of the analysis, more experienced users, familiar with these environments, try these features in order to see greater detail but the users with special needs show greater comfort with this strategy that features more simplicity of the application.

The displacement of the objects is allowed through the use of the keyboard and the mouse; this last allows greater comfort for the user, as the results show. The selection allows the choice of elements and interaction with that object, this last being referred by the group survey as an important operation in the analysis of details in this type of simulation.

The prototype in the analysis tends to be an aid element and understanding of an appropriate methodology to develop in this study and therefore provides help in video format. To do so, the user needs to click the corresponding button. The help comes in context through the action performed at the time, trying to understand and meet the needs. On its turn, the user can view all the support videos, whenever convenient, which work as a tutorial.

A friendly and engaging environment makes learning accessible and it can determine its success, an opinion shared by several authors in the field of education and information technology. Thus, this prototype arises and tries to present only the necessary elements for the user to perform the task.

The purpose of this paper is to evaluate the usability, accessibility and functionality of an educational content developed in a three dimensional environment (Figure 4). Accordingly, the above prototype was designed with the aim of promoting new challenges to the public.

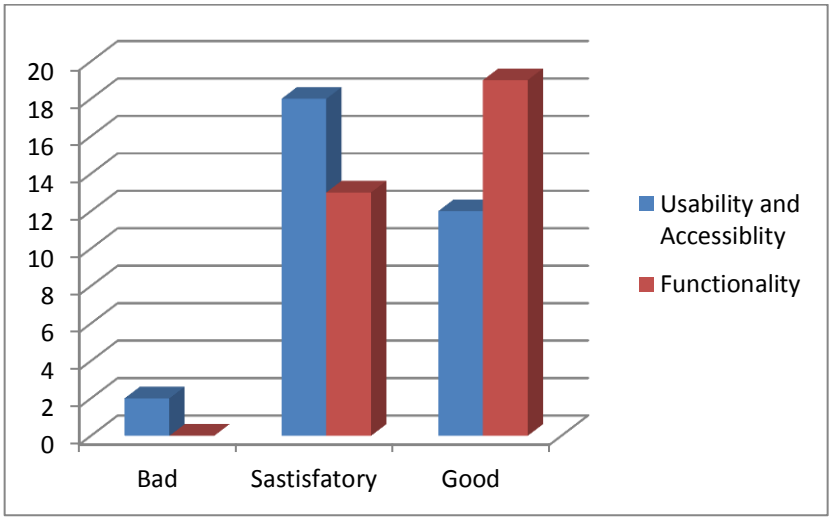


Figure 4 - Usability, accessibility and functionality of an educational

The quality of the environment exposed was evaluated as satisfactory. The navigation revealed to be an important point so, it requires greater attention. Regarding this aspect, the users sometimes showed difficulty in handling the object they wanted, but the taste by use of the model was verified.

The methodology is based in the observation of tasks, the comments about the difficulties and the analysis of the verification results. So, after identifying requirements, the prototype was developed for subsequent presentation to students. The information obtained can improve the solution presented.

The primary appeal of learn by doing simulations is that it can provide tremendously effective and engaging learning. The results obtained showed reactions to error on the part of users in the tasks required. The tutoring component provides feedback in the form of text, in order to help the students make the appropriate generalizations. Another option is a video clip provided, explaining a concept in more detail. This has been seen as an asset by users that revealed greater difficulties.

5 Conclusion

The 3D environments are present now in different situations, and one should check their behavior in the context of learning. The study comes after observing the conduct of students with this medium. The interest and commitment shown in the attempt to achieve the objectives pursued by these applications leaves its use as auxiliary tools open.

The use of a tridimensional environment allows the creation of an expectation and greater enthusiasm around the activity. The success of an application is related to the strategy and implementation used because the navigability and the interactivity appear as central elements to the user. A tool with a strong visual impact may discourage or make the understanding slower. On the other hand, the simplicity of an environment allows the user to focus and quickly assimilate the aim of the tool.

The ever-increasing realism, resulting of the advances in the area, allows users to be engaged and captivated and leaves the possibility for a greater exploitation of contents in 3D simulation environments in the learning process, especially for users with special needs, in the future.

In sum, users who have experienced the tools available in this study, specifically simulation environments for computer assembly and training simulators are satisfied with the presented environment, considered by them as pleasant and attractive, but less satisfied with the readability, clarity and consistency these for use of people with special needs. The guidance for these students in simulations should be very simple

The prototype presented highlights this desire allowing the use of a simple 3D environment. As future work, the project will continue the process of analysis and

evaluation within the school community so as to allow a final product according to their specific needs.

Also, in the future it is expected the possibility to develop other applications in different fields of knowledge for this public to facilitate learning. In this process there will be need for continued analysis of interactivity, usability and accessibility, this last aspect is important for the functionality and educational success to the student with special needs.

References

1. Colvin J., Delivering 3D, Simulation and Serious Gaming For Education and Training, 3D Visualization World Magazine, 2014
2. Youngs K., How to use 3D content in simulations for teaching and learning, Jisc Digital Festival, 2014
3. Jong T. & Joolingen W. R., Scientific discovery learning wirh computer simulations of conceptual domains, Review of Educational Research, 1998
4. Jimoytanns A., Gaming and Simulations: Concepts, Methodologies, Tools and Applications, Volume 1, New York, E.U.A., 2010
5. Glassner, A., Interactive Storytelling: techniques for 21st Century Fiction, A. K. Peters, LTD, 2004
6. Bento, J., Gonçalves, V., 3D environments in the process of teaching and learning, EDUSER: revista de educação, Vol 3(1), ESE-IPB, Portugal, 2011.
7. Bowman, A., Johnson, B., and Hodges, F., Testbed evaluation of virtual environment interaction techniques, 2001.
8. Bowman, A., Kruijff, E., J. J. L., and Poupyrev, I., An introduction to 3D user interface design, 2001.
9. Lemoine, P., Vexo, F., and Thalmann, D., Interaction techniques: 3d menus-based paradigm, In Proceedings of First Research Workshop on Augmented Virtual Reality (AVIR2003), Geneva, Switzerland, 2003.
10. Sabbadini, F., Oliveira, M., Simulação interativa visual: uma ferramenta para tomada de decisão, III Simpósio de Excelência em Gestão e Tecnologia, UFRJ, Brasil, 2006.
11. Galland S., Gaud N., Demange J., Koukam A., Multilevel Model of the 3D Virtual Environment for Crowd Simulation in Buildings, The 5th International Conference on Ambient Systems, Networks and Technologies, Belgium, 2014.
12. D. Cohen-Or, Y.L. Chrysanthou, C.T. Silva, F. Durand, A survey of visibility for walkthrough applications, IEEE Transactions on Visualization and Computer Graphics, 2003.
13. Eve E., Koo S., Alshihri A., Cormier J., Kozhenikov M., Donoff R., Karimbux N., Performance of Dental Students Versus Prosthodontics Residents on a 3D Immersive Haptic Simulator, 2013.

14. Reiners T., Teräs H., Chang V., Wood L., Gregory S., Gibson D., Petter N., Teräs M., Authentic, immersive, and emotional experience in virtual learning environments: The fear of dying as an important learning experience in a simulation, Teaching and Learning Forum, Australia, 2014
15. Brookes, S. & Moseley, A., Authentic contextual games for learning. In N. Whitton & A. Moseley (Eds.), Using games to enhance learning and teaching: A beginner's guide, U.S.A., 2012.
16. Diewald S., Möller A., Roalter L., Kranz M., Simulation and Virtual Prototyping of Tangible User Interfaces, Library Cornell University, 2014.
17. Guralmick D.A. & Levy C., Educational Simulations: Learning from the Past and Ensuring Success in the Future, Volume 1, New York, E.U.A., 2010.

PAPERS IN ALPHABETICAL ORDER

Analysis and Evaluation of gesture recognition using LeapMotion 83

A Multi-player Approach in Serious Games: Testing Pedestrian Fire Evacuation Scenarios 120

A Review of recent progress in multi document summarization 48

A Survey of Merging Decision Trees Data Mining Approaches 36

Characterizing Developers' Rework on GitHub Open Source Projects 70

Context-based learning games for children with cerebral palsy: a prototype 25

Online forums in Higher Education: empowering female participation 13

Procedural Generation of Maps and Narrative Inclusion for Video Games 106

Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm 95

Text Mining Scientific Articles using the R Language 60

3D Simulation Environment: Education and Training 132

AUTHORS IN ALPHABETICAL ORDER

António Coelho	106
Carlos Gulo	60, 70
Elis Silva	25
Eugénio Oliveira	48, 120
Hugo Barbosa	132
João Emilio Almeida	120
João Ulisses	106
Jorge Silva	25
Luciano Moreira	13
Marcos Oliveira	120
Nelson Pereira	120
Pedro Leitão	83
Pedro Strecht	36
Ricardo Gonçalves and	106
Rosaldo Rossetti	120
Shazia Tabassum	48
Telmo Morais	95
Thiago Rúbio	70, 60


9 789727 521739 >